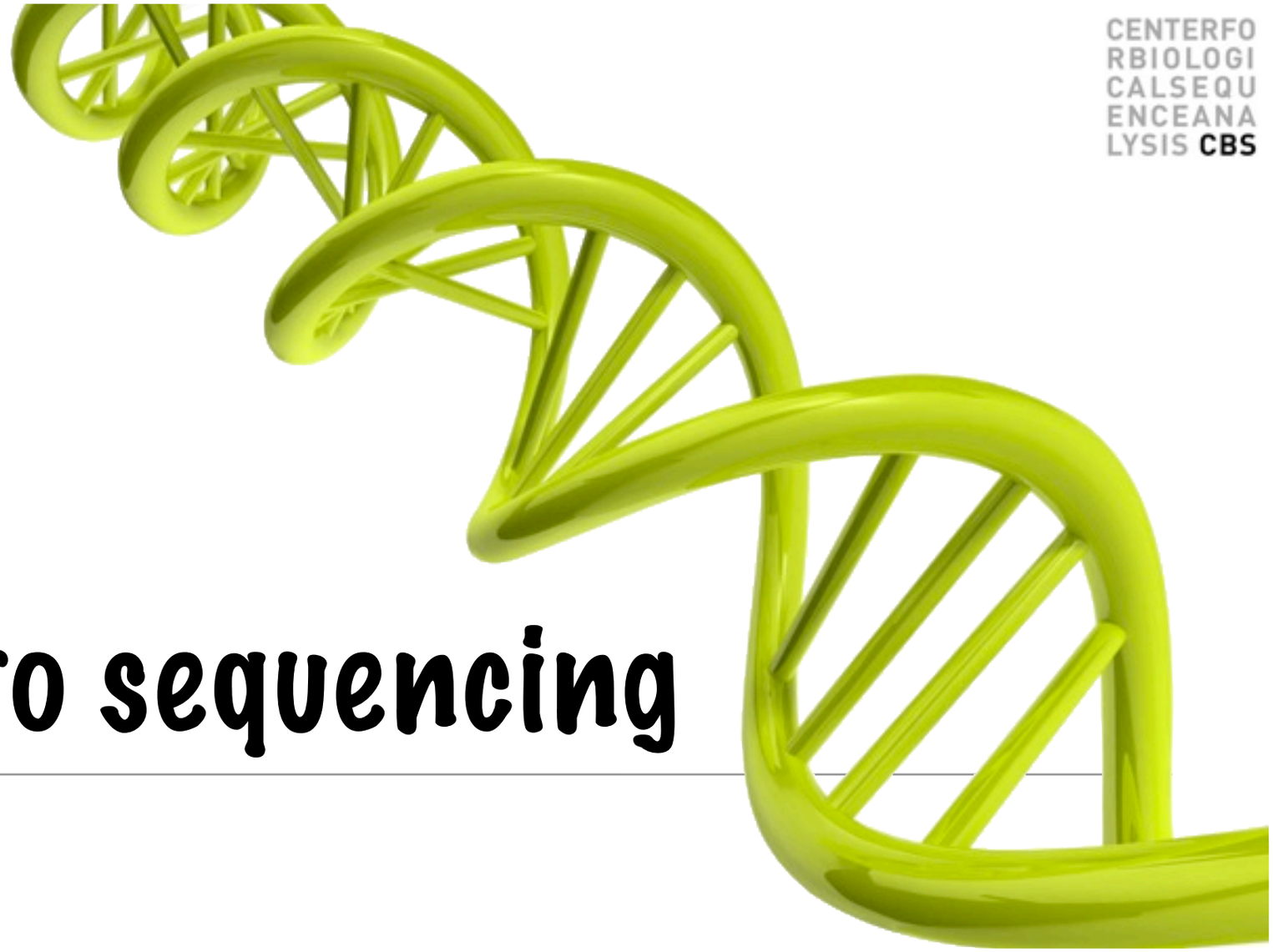


CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

# Introduction to sequencing

**Kirstine Belling**

Introduction to Systems Biology

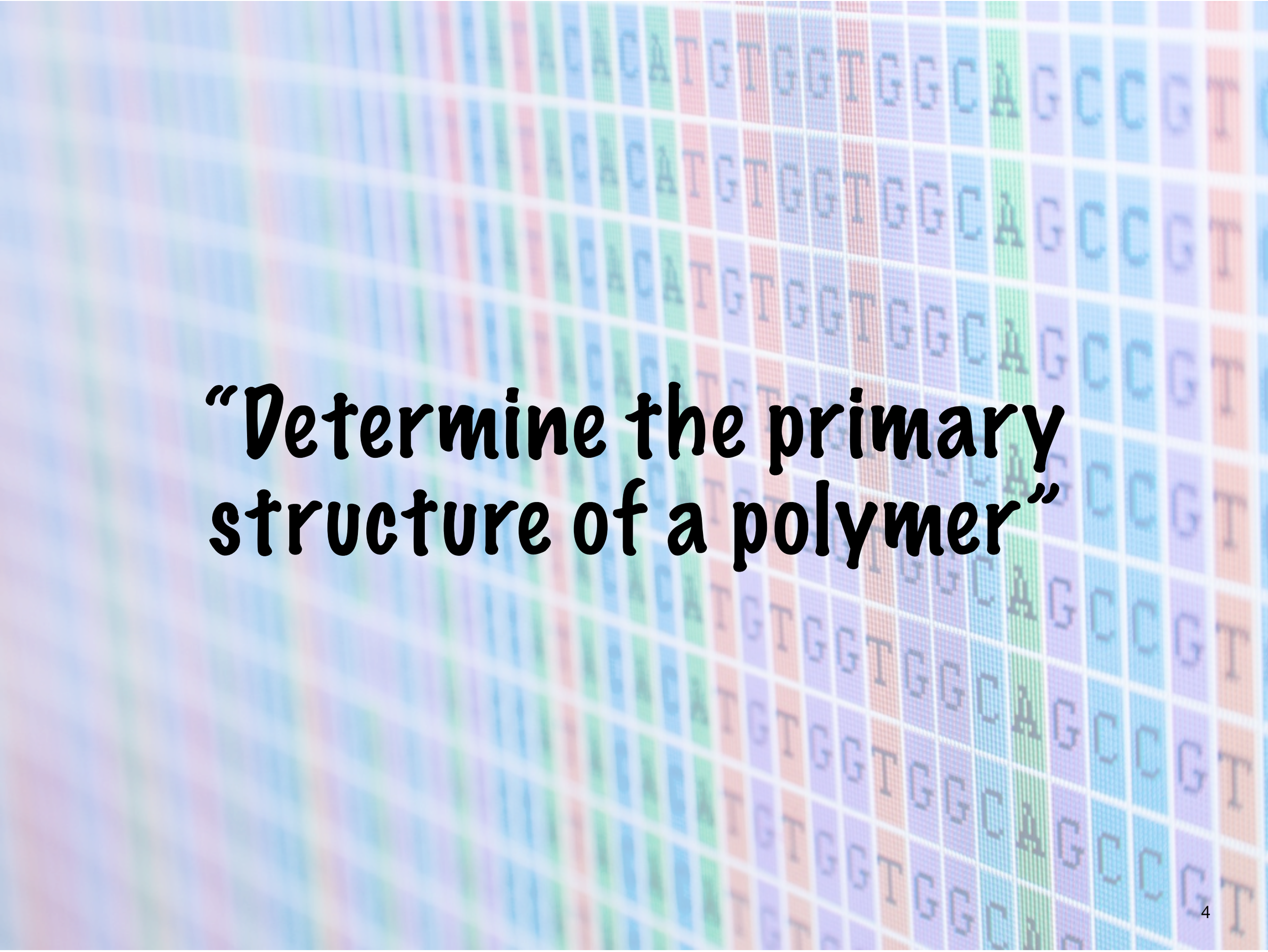
February 11th 2014

# Agenda

- Sequencing
- Sequencing technology
- Data files and formats
- Analysis
- Introduction to exercise

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

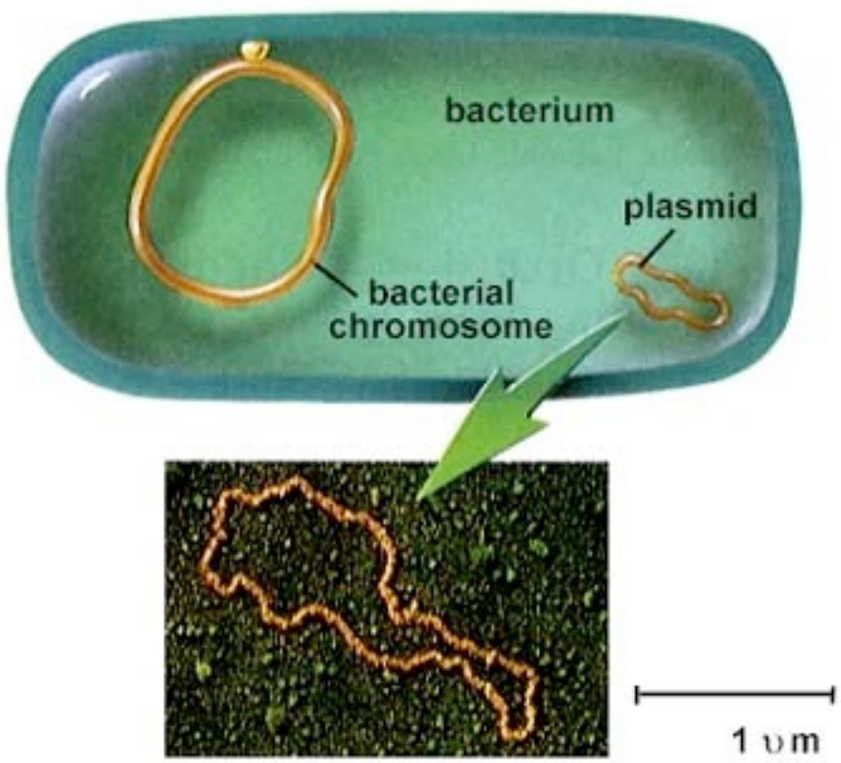
# Sequencing



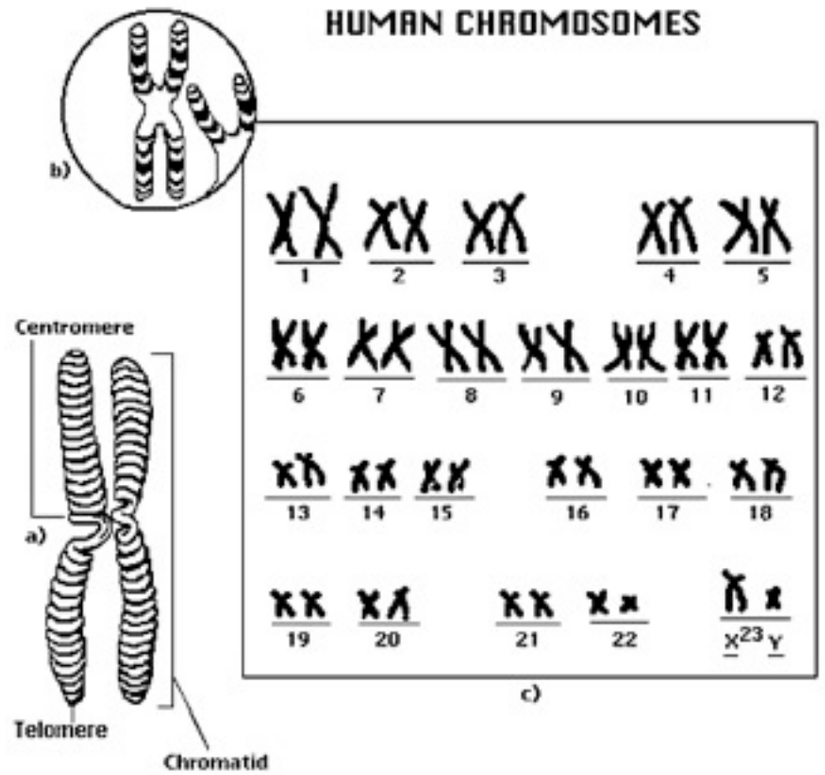
"Determine the primary structure of a polymer"

# Genome sequencing

- Determine the nucleotide sequence of all chromosomes in a genome
  - Prokaryotes: One chromosome
  - Eukaryotes: Sequence all autosome and sex chromosomes




http://biology.kenyon.edu/courses/biol114/Chap01/bact-chrom.jpg



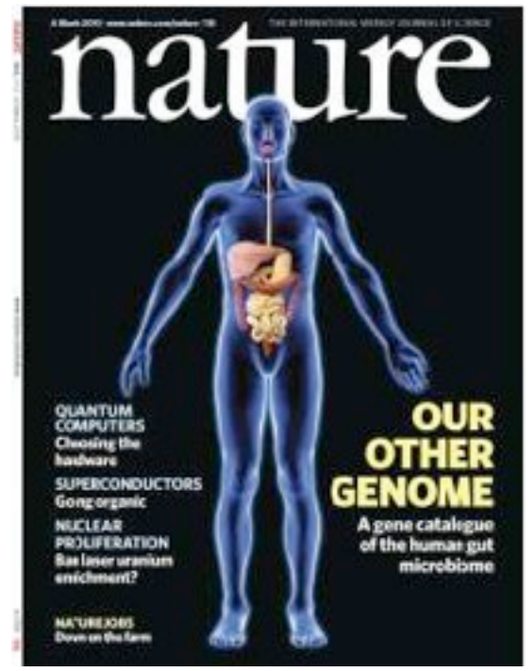


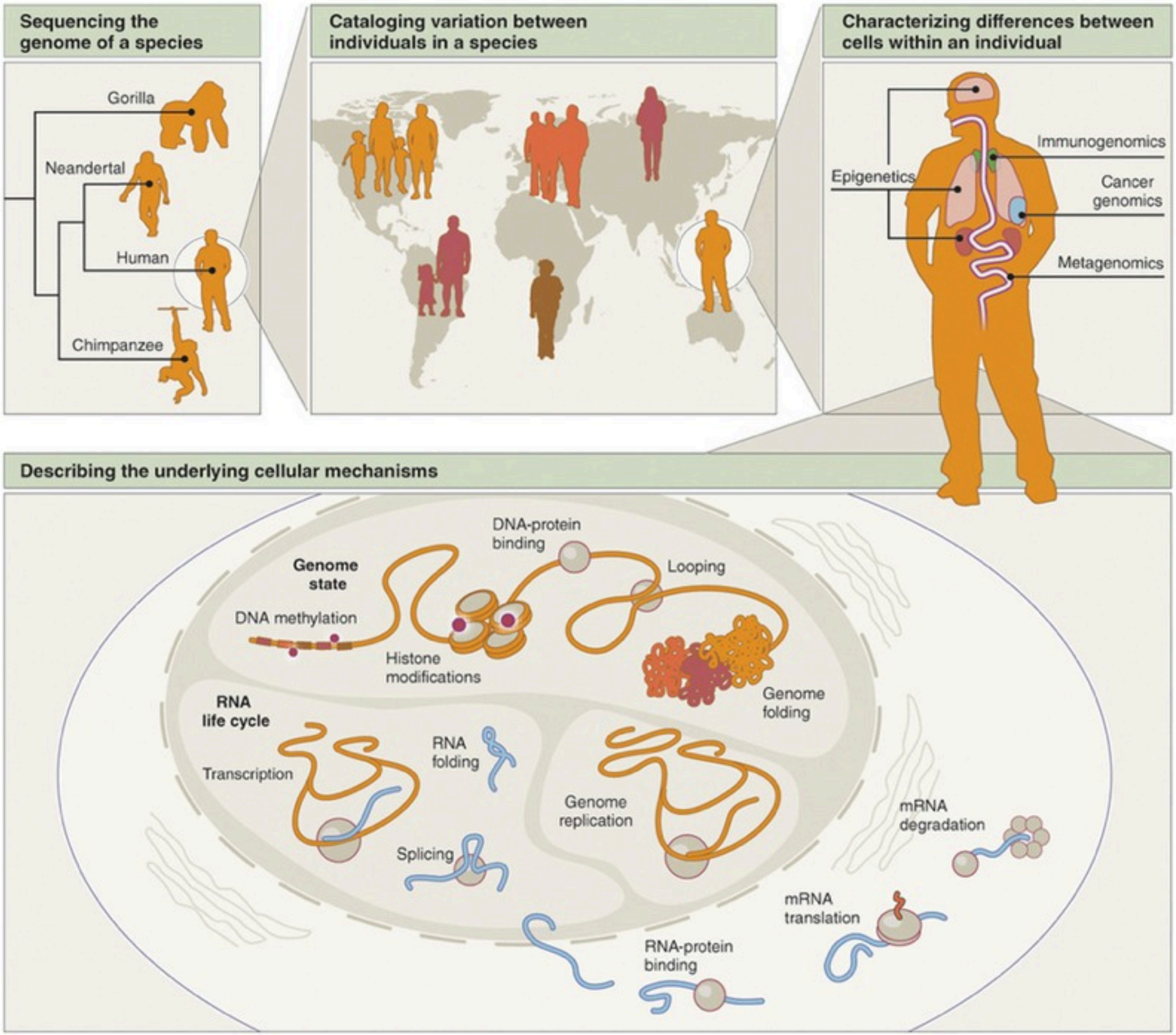



5

# What is sequencing useful for?

- Sequence organisms
- Genome re-sequencing
- Determine the DNA sequence of prehistoric humans
- Metagenomics
- Cancer genomics
- Exon sequencing (~2% af genomet)
- *...*




Shendure & Aiden, 2012

# Genome Denmark

## Et dansk reference genom

**Hans Eiberg**

# research interests

## Resourcecenter for linkage analysis (RCLINK)
## Copenhagen family bank

**projects:**

- Resource center for linkage analysis (850 families with at least 4 children)
- **Family structure**
- 50 families with more than 6 children tested for more than 400 markers.
- Serum,plasma,erythrocytes from all. B-lymphocytes,DNA,mRNA from 300 families
- Psoriasis (50 informative families)
- Asthma & hay fever (120 informative families with at least 4 children)
- Enuresis 18 families
- Migraine 30 families
- Common traits (such as eye colour, hair colour etc.)
- Eye diseases (cataract, opticus atrophy, Groenouw I)
- Psychiatric diseases (depressions,schizofrenia)
- NIDDM

Populær artikel om projektet: Bio-nyt nr 95 side 1-22. Biobanker 1. del. 1996.

# Resourcecenter for linkage analysis (RCLINK)
# Copenhagen family bank

## family structure

Families total 848 in family bank from 1973

---

650 fam. with 4 children
108 fam. with 5 children
48 fam. with 6 children
15 fam. with 7 children
12 fam. with 8 children
5 fam. with 9 children
6 fam. with 10 children
1 fam. with 11 children
2 fam. with 12 children
1 fam. with 17 children

Grandparents: 702 (14% alive march 2000)
Parents: 1680
Families collected in 1973-1974, a few later.

Most grandparents born from 1880-1920
Most parents born from 1920-1940
Most children born from 1950-1968

All persons(about 6000) tested for 80 classical markers (ABO,HLA,PGM,HP....)
Information from ca. 600 families, about common traits and diseases (eye color, psoriasis, astma, migraine, enuresis, dyslexia, ......).

## Materials

Serum, plasma, saliva, erythrocytes, thrombocytes, leucocytes in liquid nitrogen.
From 350 families (large): DNA, RNA, transformed B-cells.
Grandparends also: skin biopsy, hair.

UNIVERSITETETS ARVEBIOLOGISKE INSTITUT

*Bestyrer: professor dr. med. Jan Mohr*

TAGENSVEJ 14, 2200 KØBENHAVN N - TELEFON (01) 39 33 73

Hr. og Fru,

20. 2. 73.

[redacted]

Arveligheds-undersøgelse

Vi vil gerne besøge Dem

torsdag d. 1. marts 1973 mellem kl. 18 og 19.

for at få en almindelig blodprøve og en spytprøve fra hele familien til brug for en kortlægning af den normale arvegang for blod- og enzymtyper.

Besøget varer kun nogle få minutter pr. person, og vi håber, at De vil samarbejde med os alene for den gode sags skyld.

Desværre er vi ikke i stand til egentlig at honorere Deres medvirken, men vi udbetaler 5 kroner til hver person, der deltager i undersøgelsen.

Det er vort mål at samle prøver fra omkring 1000 familier, og De forstår sikkert, at det har betydning, at vi undgår at køre forgæves, hvorfor vi venligst beder Dem bekræfte Deres deltagelse ved straks at returnere svarbrevet i den medfølgende konvolut.

Der er forklaret lidt mere om undersøgelsens art og formål i den vedlagte skrivelse.

Med venlig hilsen

Jan Mohr

Jan Mohr
professor, dr.med.

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

# Sequencing technology

# Sequence technologies

- Illumina
- SOLID
- 454
- Ion Torrent
- Pacific BioSciences
- Oxford Nanopore

# Sequence technologies

- Illumina
- SOLID
- 454
- Ion Torrent
- Pacific BioSciences
- Oxford Nanopore

# Sequence technologies

| Platform | Library/ template preparation | NGS chemistry | Read length (bases) | Run time (days) | Gb per run | Machine cost (US$) | Pros | Cons | Biological applications | Refs |
|---|---|---|---|---|---|---|---|---|---|---|
| Roche/454's GS FLX Titanium | Frag, MP/ emPCR | PS | 330* | 0.35 | 0.45 | 500,000 | Longer reads improve mapping in repetitive regions; fast run times | High reagent cost; high error rates in homo-polymer repeats | Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics | D. Muzny, pers. comm. |
| Illumina/ Solexa's $GA_{II}$ | Frag, MP/ solid-phase | RTs | 75 or 100 | 4‡, 9§ | 18‡, 35§ | 540,000 | Currently the most widely used platform in the field | Low multiplexing capability of samples | Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics | D. Muzny, pers. comm. |
| Life/APG's SOLiD 3 | Frag, MP/ emPCR | Cleavable probe SBL | 50 | 7‡, 14§ | 30‡, 50§ | 595,000 | Two-base encoding provides inherent error correction | Long run times | Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics | D. Muzny, pers. comm. |
| Polonator G.007 | MP only/ emPCR | Non-cleavable probe SBL | 26 | 5§ | 12§ | 170,000 | Least expensive platform; open source to adapt alternative NGS chemistries | Users are required to maintain and quality control reagents; shortest NGS read lengths | Bacterial genome resequencing for variant discovery | J. Edwards, pers. comm. |
| Helicos BioSciences HeliScope | Frag, MP/ single molecule | RTs | 32* | 8‡ | 37‡ | 999,000 | Non-bias representation of templates for genome and seq-based applications | High error rates compared with other reversible terminator chemistries | Seq-based methods | 91 |
| Pacific Biosciences (target release: 2010) | Frag only/ single molecule | Real-time | 964* | N/A | N/A | N/A | Has the greatest potential for reads exceeding 1 kb | Highest error rates compared with other NGS chemistries | Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks | S. Turner, pers. comm. |

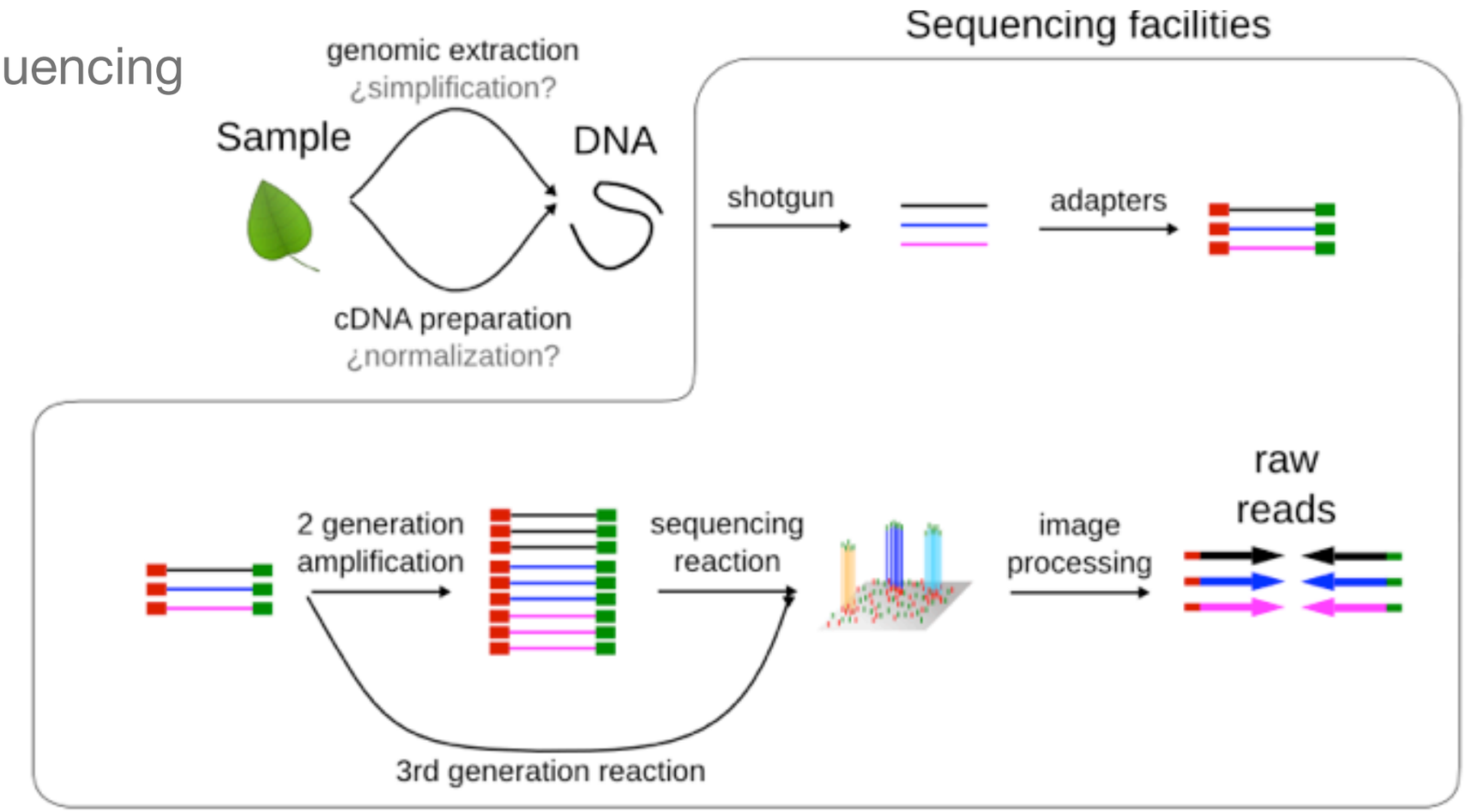
*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

# Experimental steps in sequencing

1. Make a library of DNA molecules

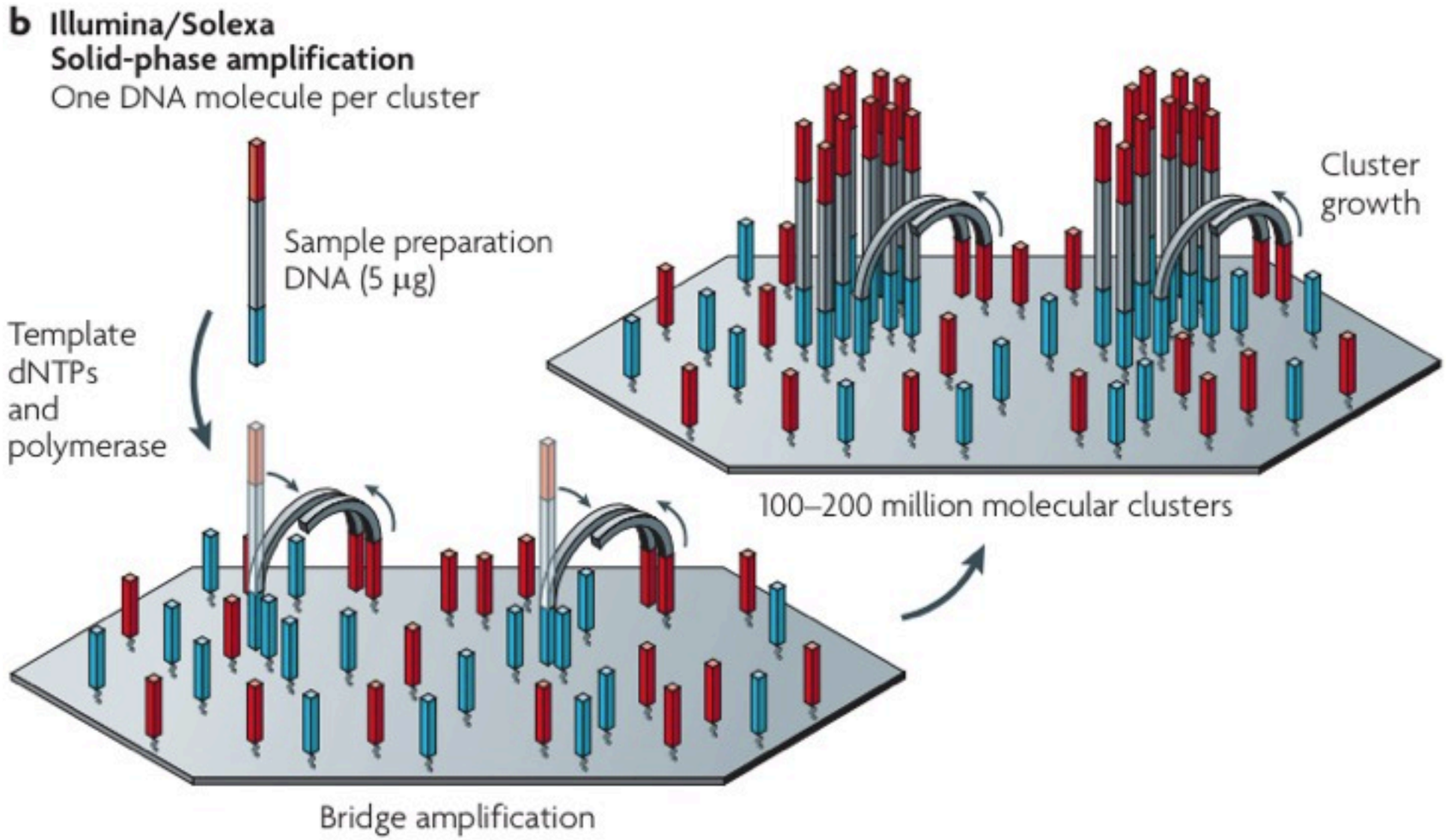
2. Amplify DNA by Polymerase chain reaction (PCR)

3. Massive parallel sequencing

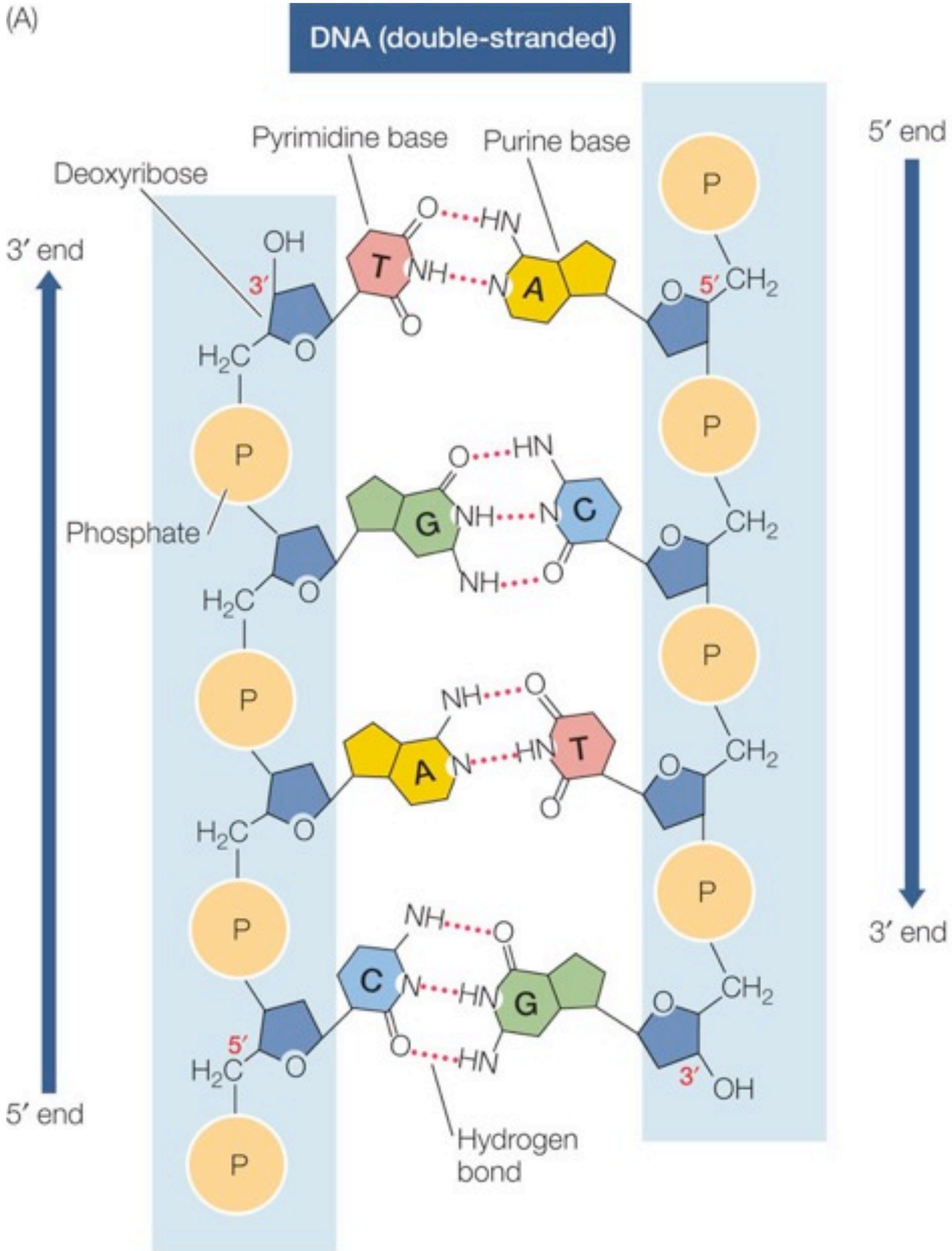



http://bioinf.comav.upv.es/_images/ngs_map.png

# Amplify and immobilize



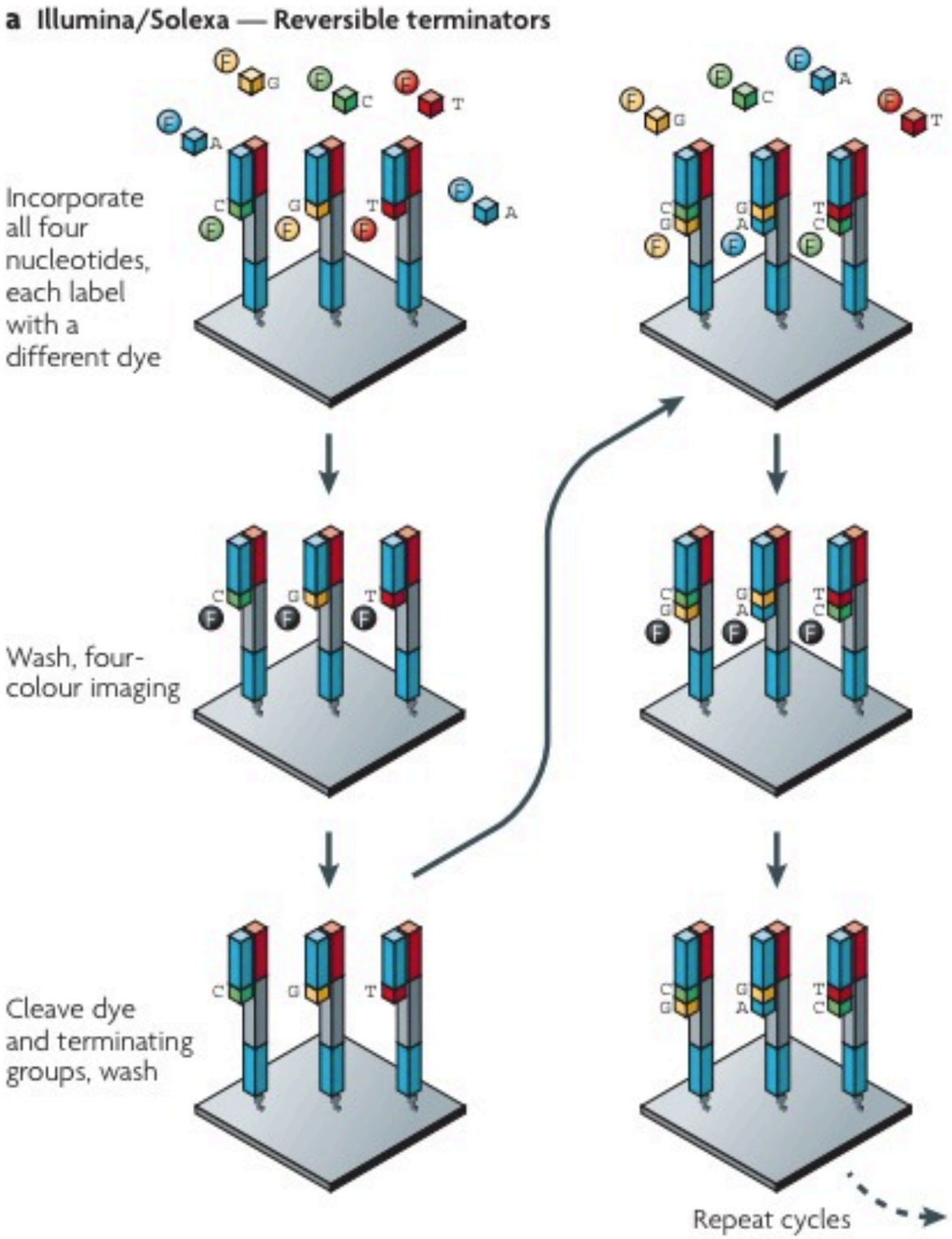

Metzker, 2010




In DNA, the bases are attached to deoxyribose, and the base thymine (T) is found instead of uracil. Hydrogen bonds between purines and pyrimidines hold the two strands of DNA together.

***LIFE 10e*, Figure 4.4**

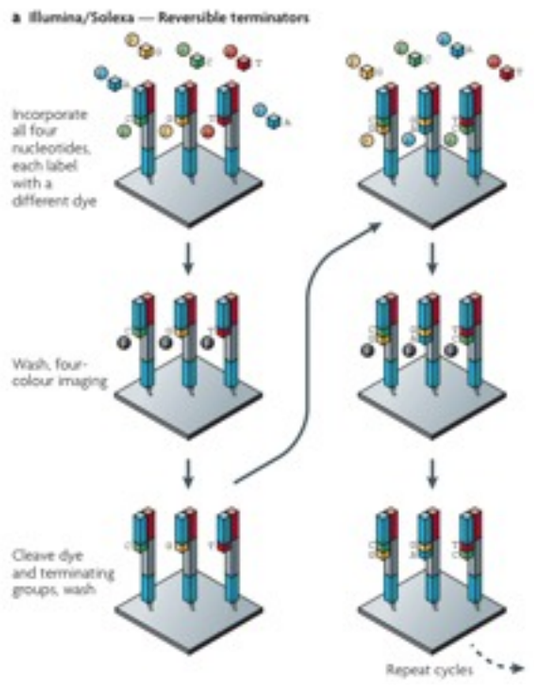
# Fluorescence detection

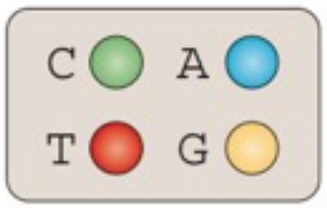



Metzker, 2010

http://www.youtube.com/watch?v=77r5p8IBwJk

# Determine the DNA sequence

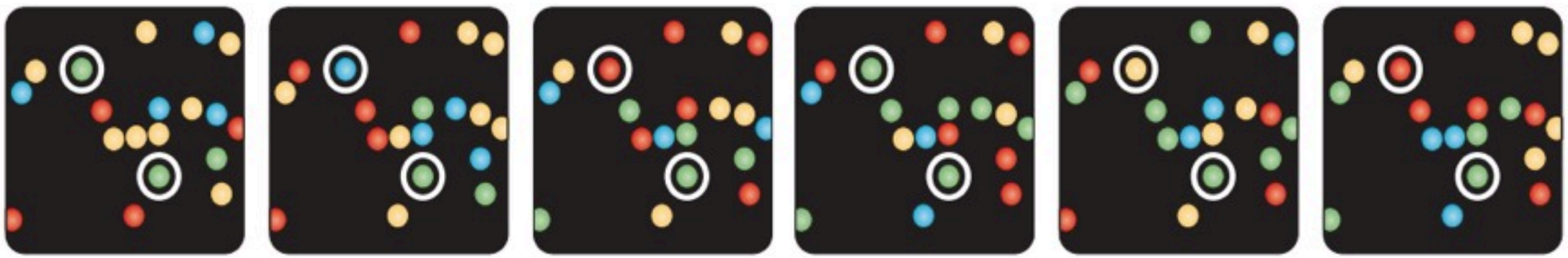


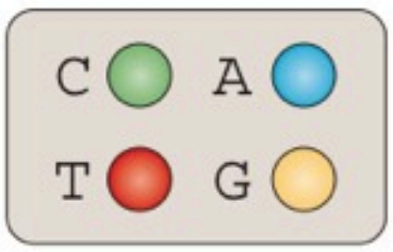




Illumina 1_____

Illumina 2_____

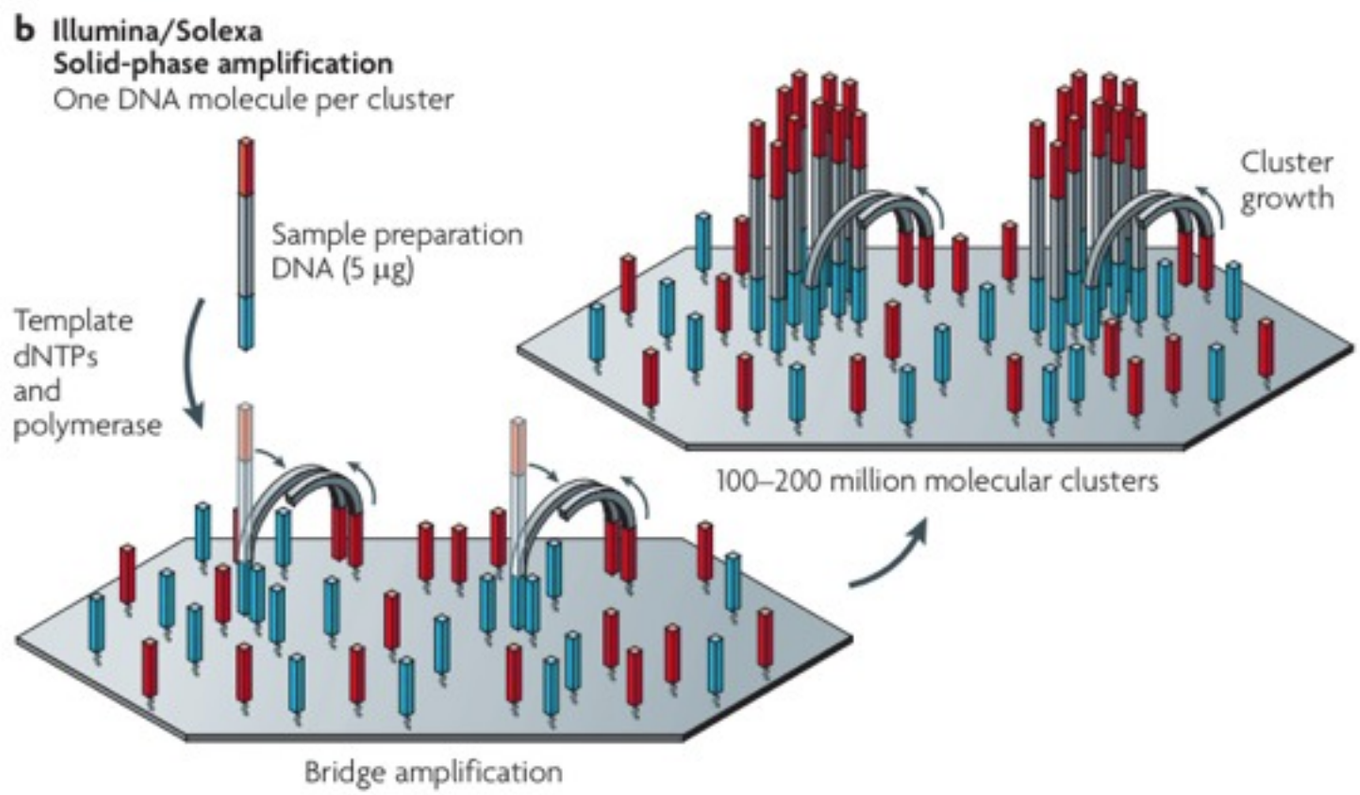
# Determine the DNA sequence



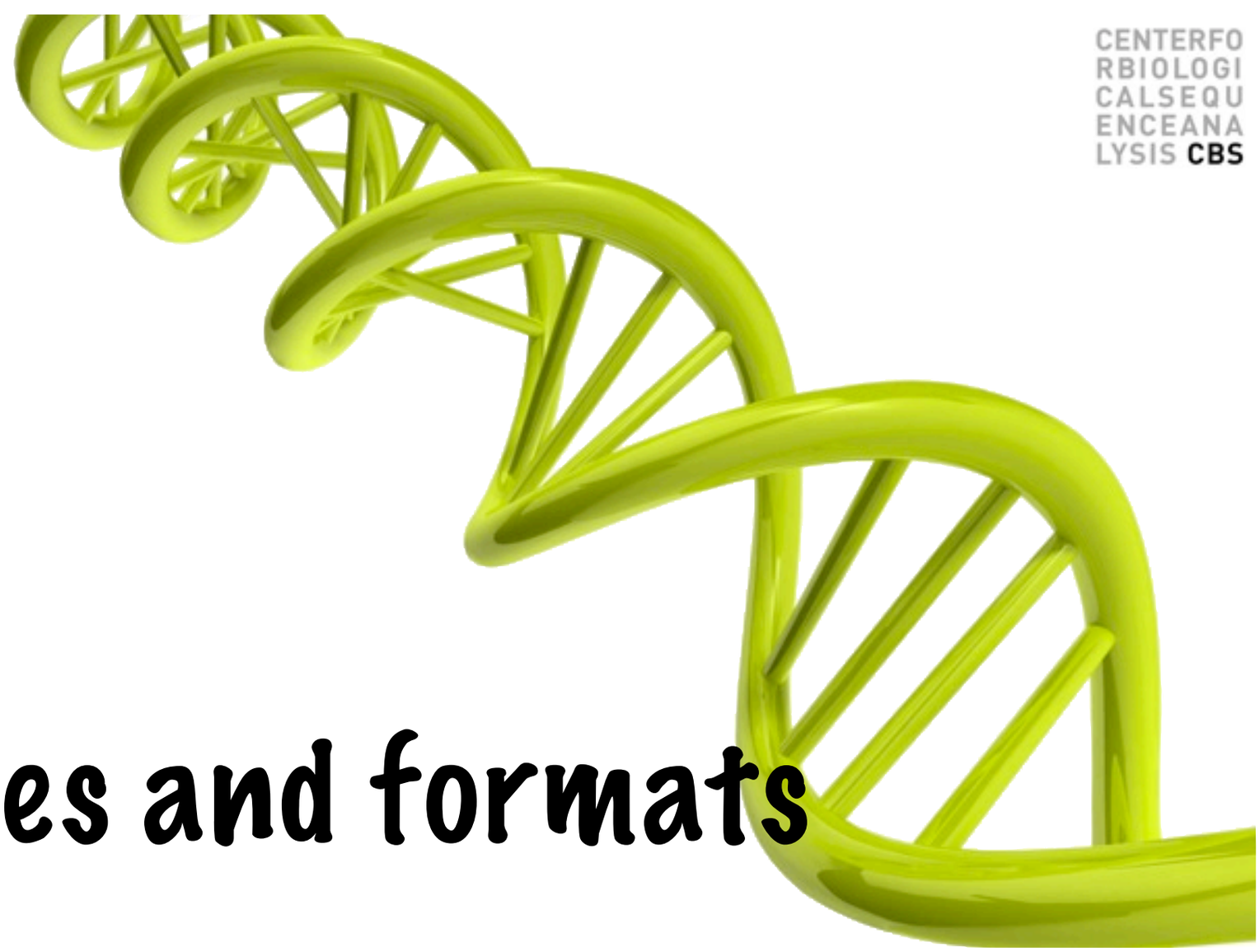



Top: CATCGT
Bottom: CCCCCC

# Reads




- **Single-end reads**
  - Fragments sequenced from one end
- **Paired-end reads**
  - Fragments sequenced from both ends



# Data files and formats

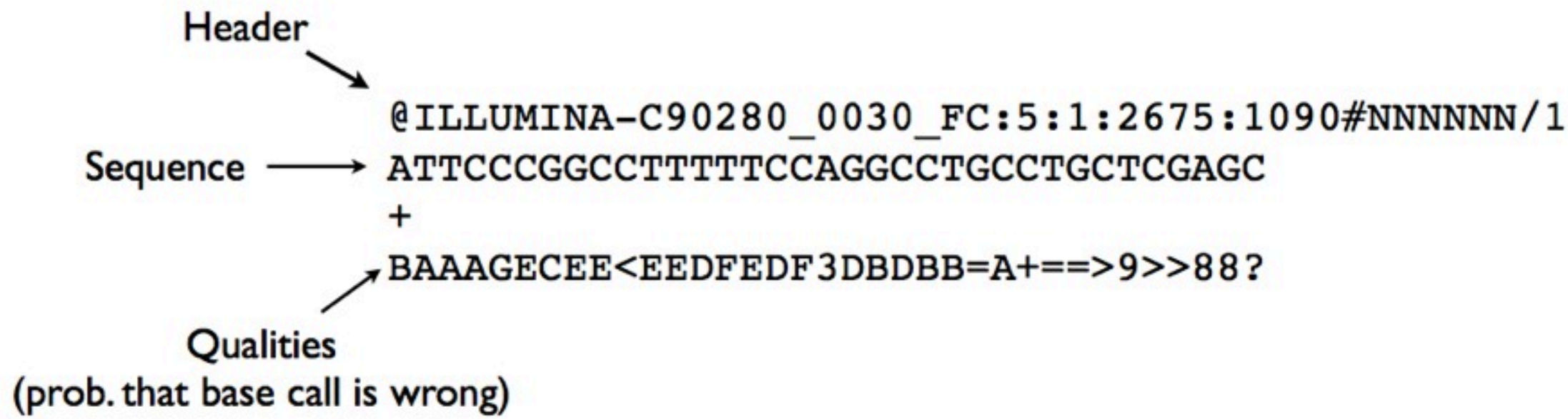
# Data files and formats

- Paired-end sequencing
  - Two files generated for each sample
  - Each fastq file contains ~95,000,000 lines

```
interaction[belling]:/home/panfs/cbs/projects/breastcancer/belling/RNA_seq_data> ll
total 46039552
-rw-r----- 1 belling cdrom 4883517198 Feb 17 16:37 101208_I117_FC80AKJABXX_L3_HUMfrwTARAAPEI-5_1.fq
-rw-r----- 1 belling cdrom 4883517198 Feb 17 16:38 101208_I117_FC80AKJABXX_L3_HUMfrwTARAAPEI-5_2.fq
-rw-r----- 1 belling cdrom 5431922303 Feb 17 16:39 101208_I117_FC80AKJABXX_L3_HUMfrwTBRAAPEI-6_1.fq
-rw-r----- 1 belling cdrom 5431922303 Feb 17 16:39 101208_I117_FC80AKJABXX_L3_HUMfrwTBRAAPEI-6_2.fq
-rw-r----- 1 belling cdrom 5224871083 Feb 17 16:40 101208_I117_FC80AKJABXX_L3_HUMfrwTCRAAPEI-7_1.fq
-rw-r----- 1 belling cdrom 5224871083 Feb 17 16:40 101208_I117_FC80AKJABXX_L3_HUMfrwTCRAAPEI-7_2.fq
-rw-r----- 1 belling cdrom 5391641213 Feb 17 16:41 101208_I117_FC80AKJABXX_L3_HUMfrwTDRAAPEI-8_1.fq
-rw-r----- 1 belling cdrom 5391641213 Feb 17 16:41 101208_I117_FC80AKJABXX_L3_HUMfrwTDRAAPEI-8_2.fq
```

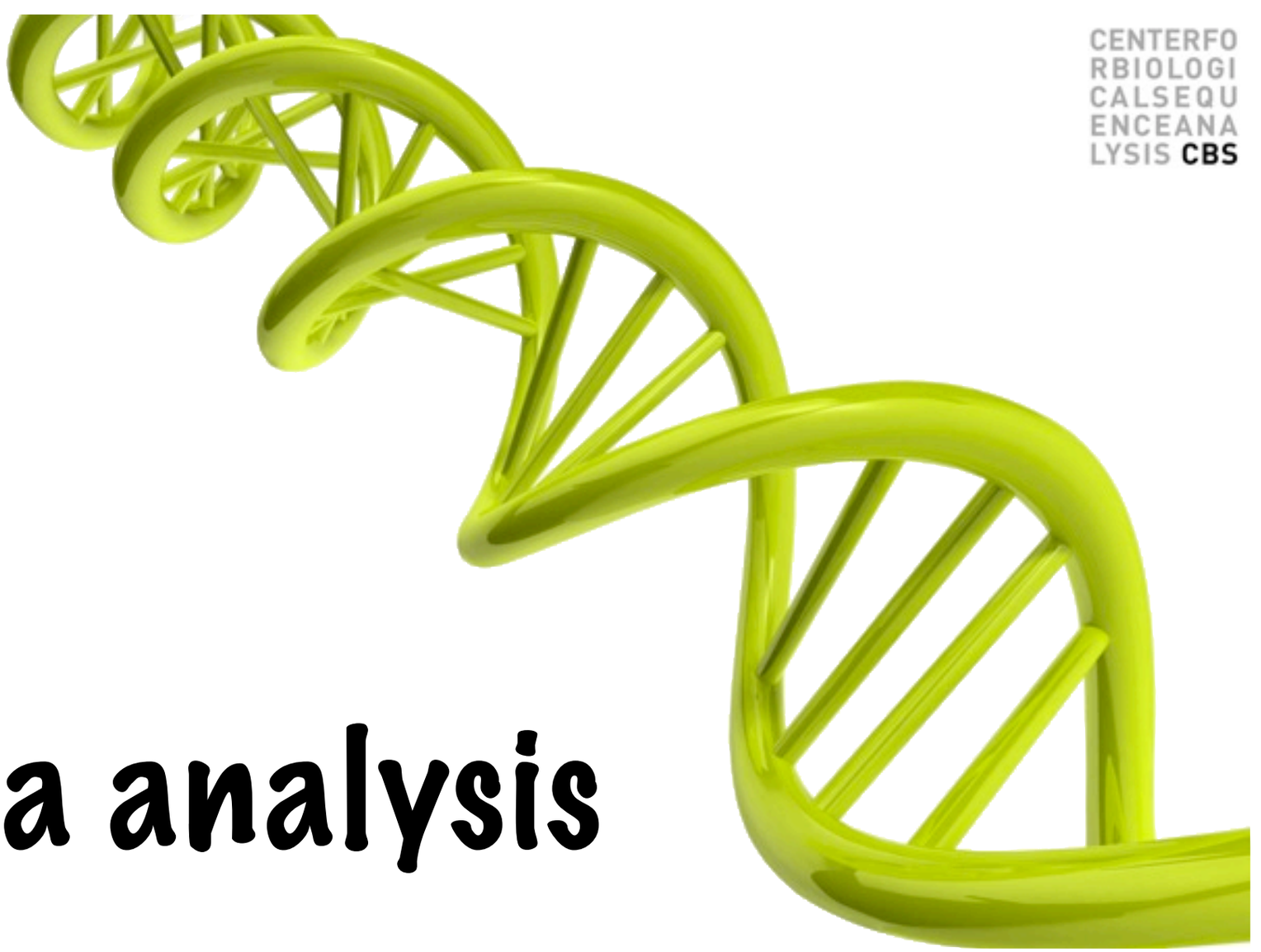
# Fastq format

1. Sequence identifier starting with an @
2. Raw sequence letters
3. +
4. Quality values - one value every nucleotide in read



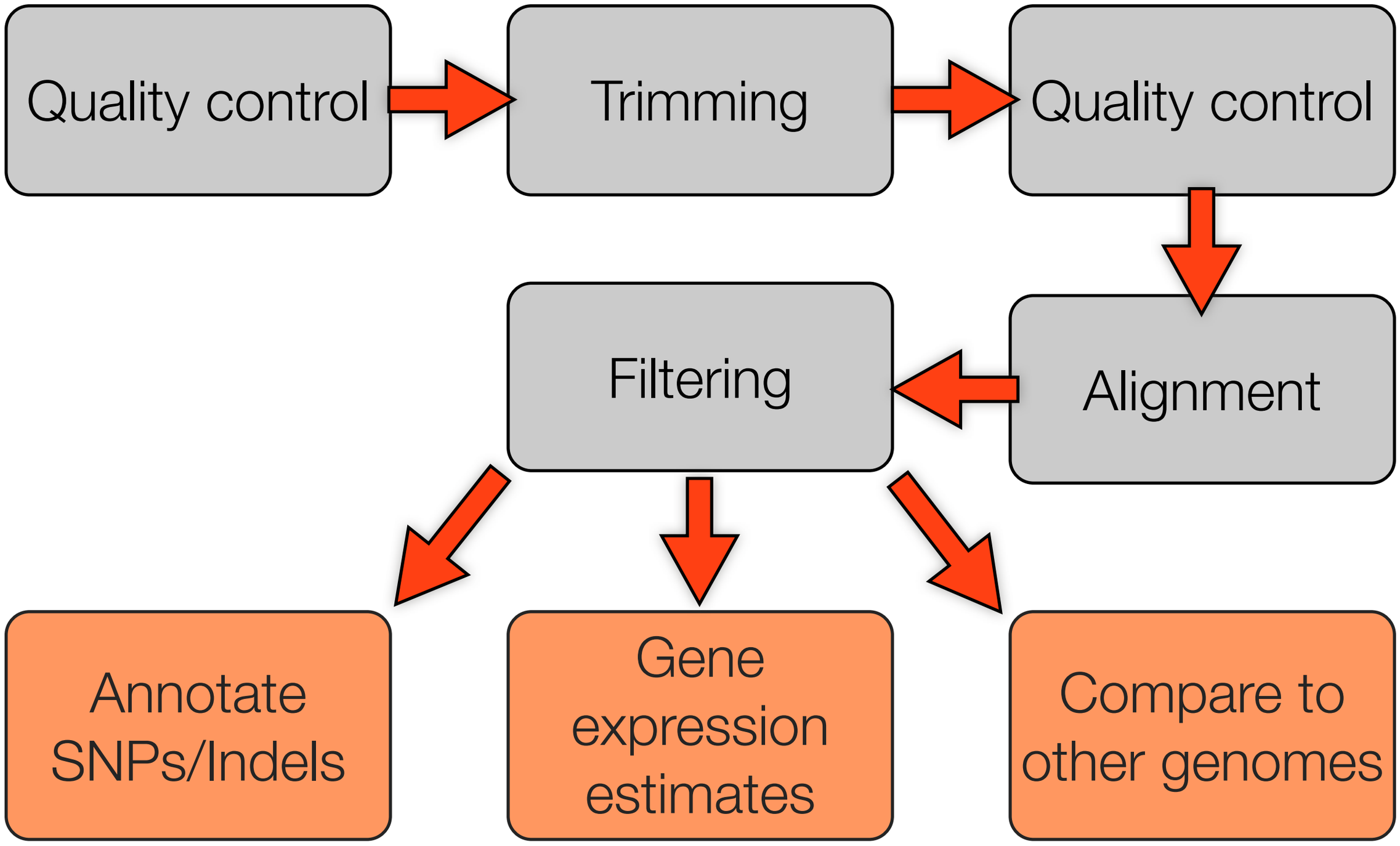
# Important numbers

- Coverage
    - How many nucleotides are covered by at least one read
- Depth
    - Depth of coverage
    - The number of reads covering a nucleotide
        A. Average
        B. At a given position



# Data analysis

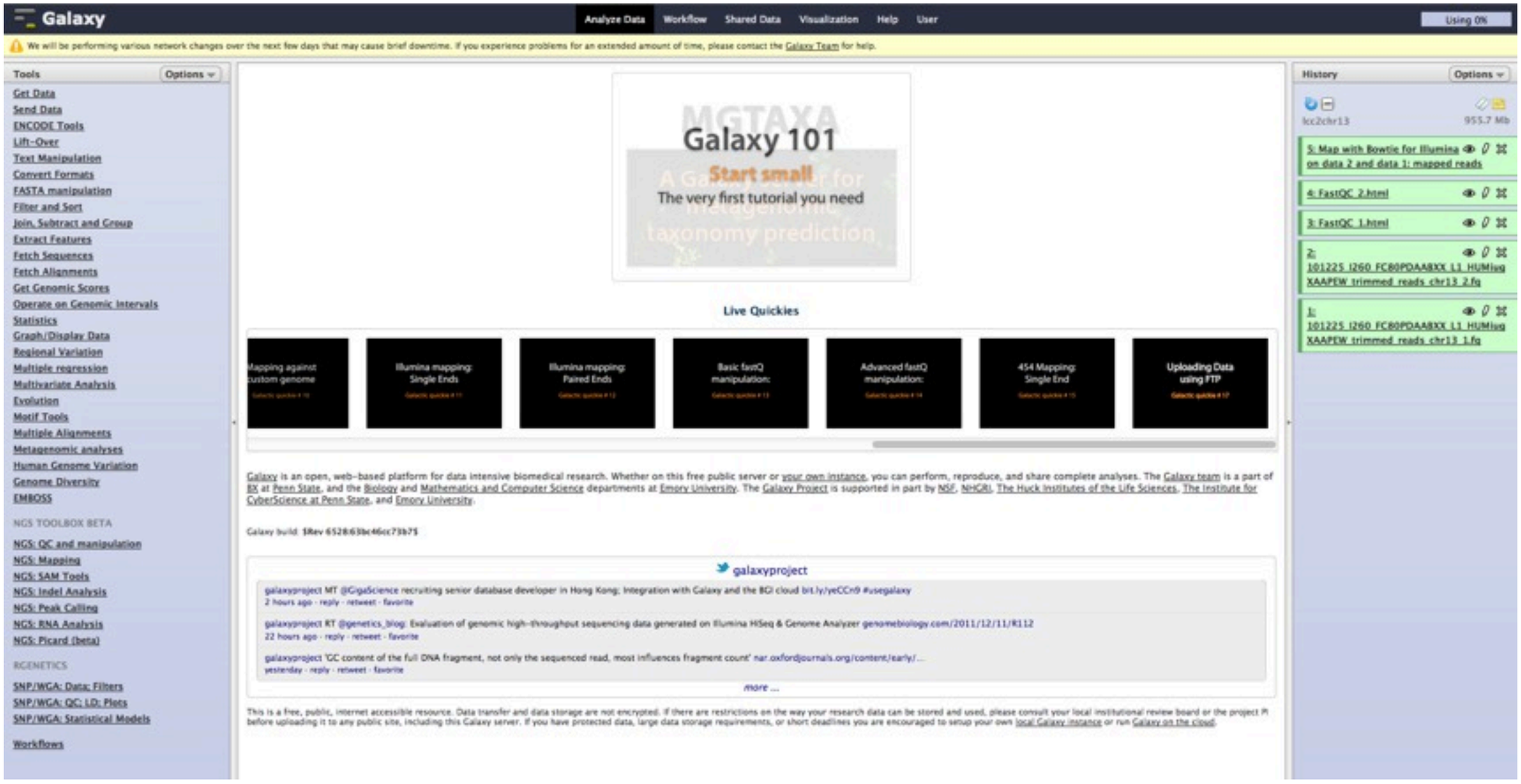
# Typical workflow of sequencing analysis



# Data analysis

1. Galaxy is one way to do it

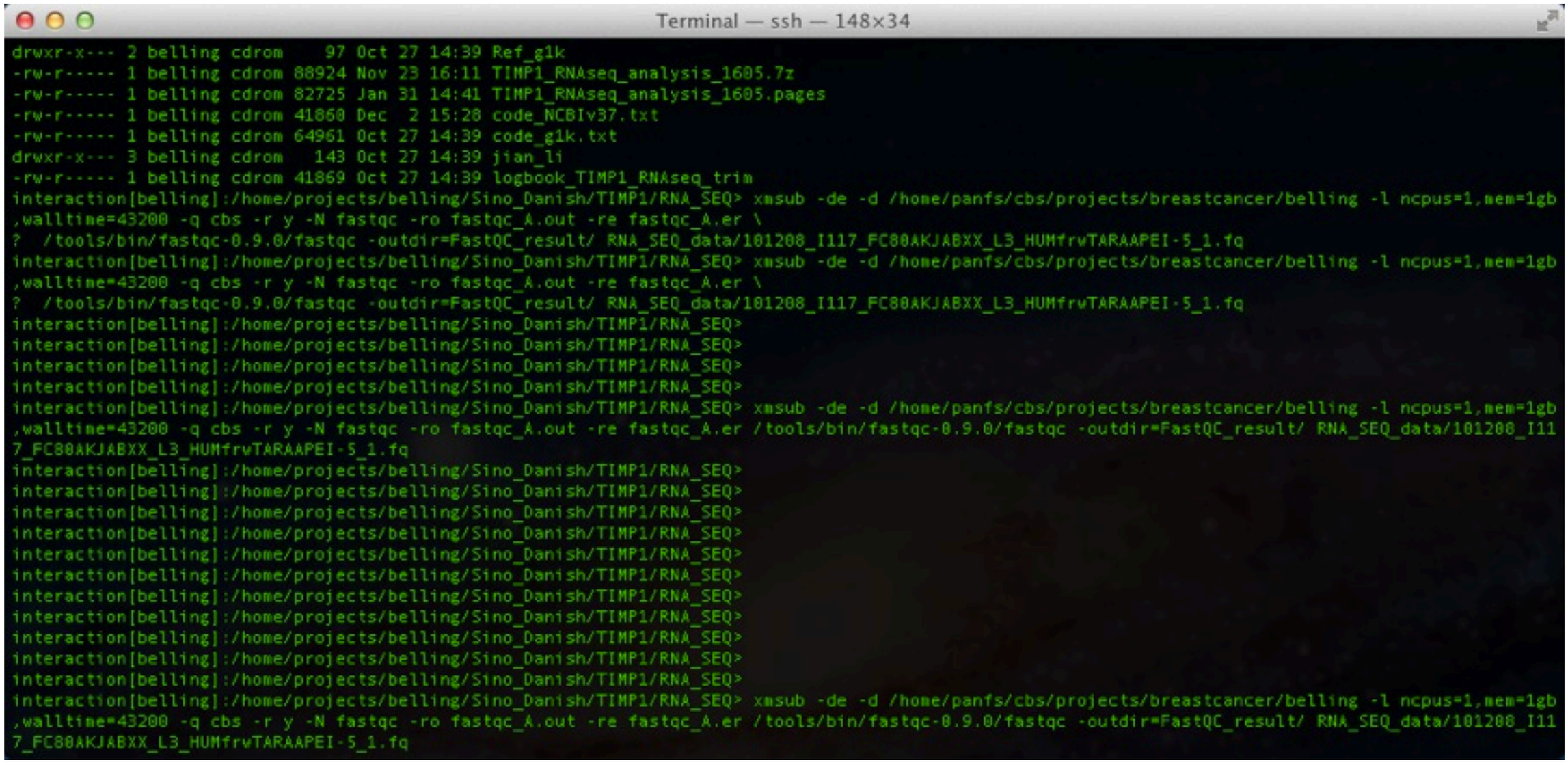
2. CBS usually we run programs via a terminal



# Data analysis

1. Galaxy is one way to do it

2. CBS usually we run programs via a terminal



# Quality control - FastQC

- Various quality parameters

**Thu 17 Feb 2011**

**101208_I117_FC80AKJABXX_L3_HUMfrwTARAAPEI-5_1.fq**

## Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

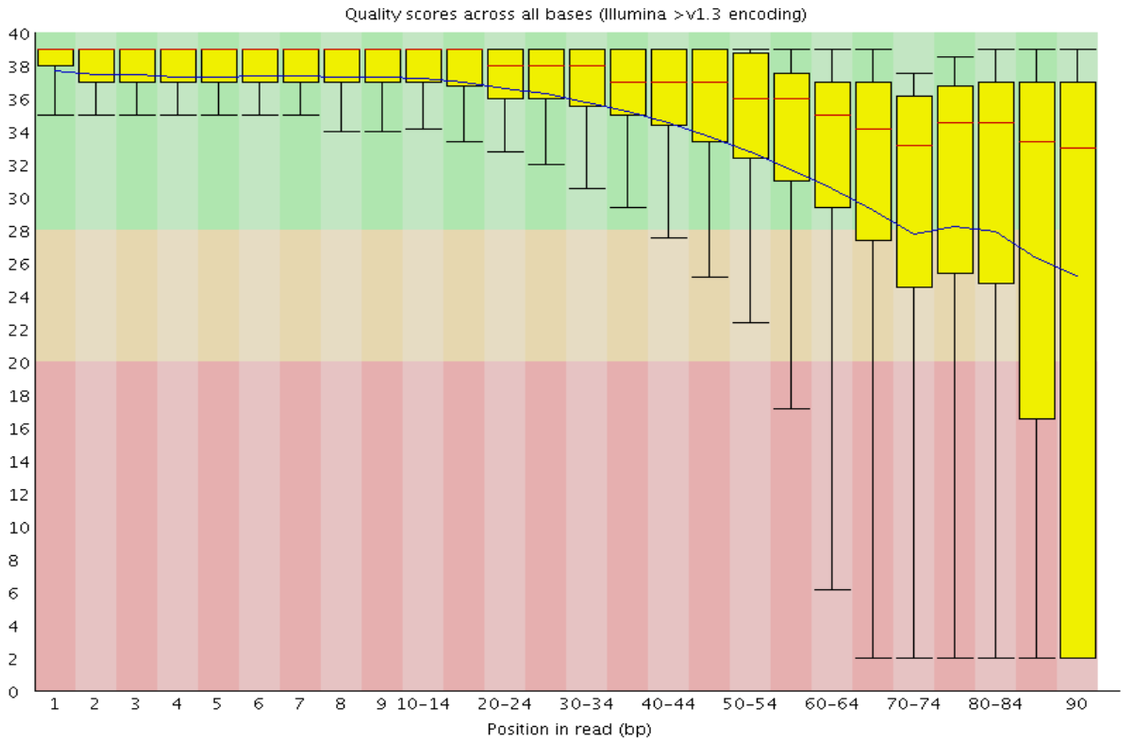
## Basic Statistics

| Measure | Value |
|---|---|
| Filename | 101208_I117_FC80AKJABXX_L3_HUMfrwTARAAPEI-5_1.fq |
| File type | Conventional base calls |
| Total Sequences | 21822999 |
| Sequence length | 90 |
| %GC | 47 |

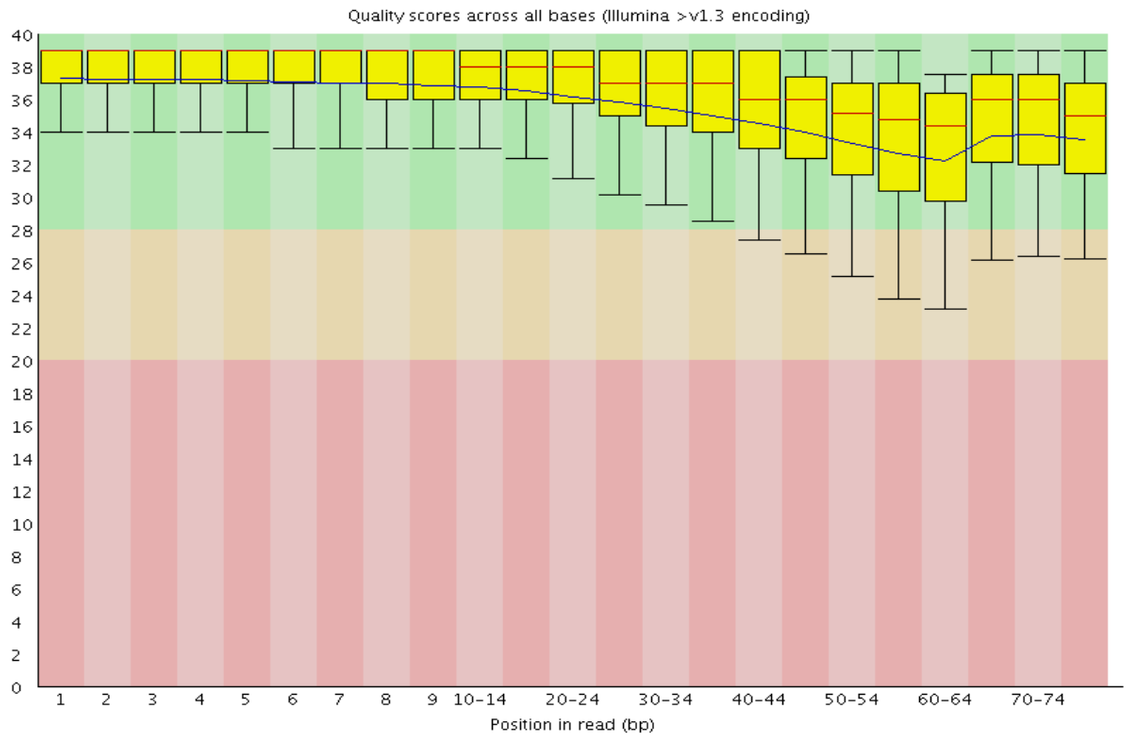
# Quality control

- Per base sequence quality

Before trimming

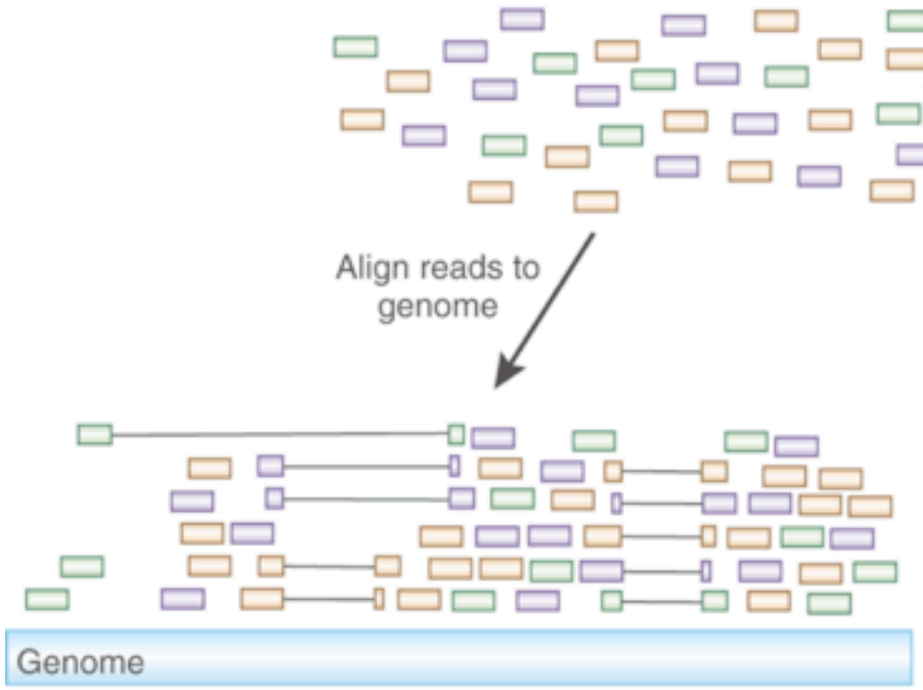


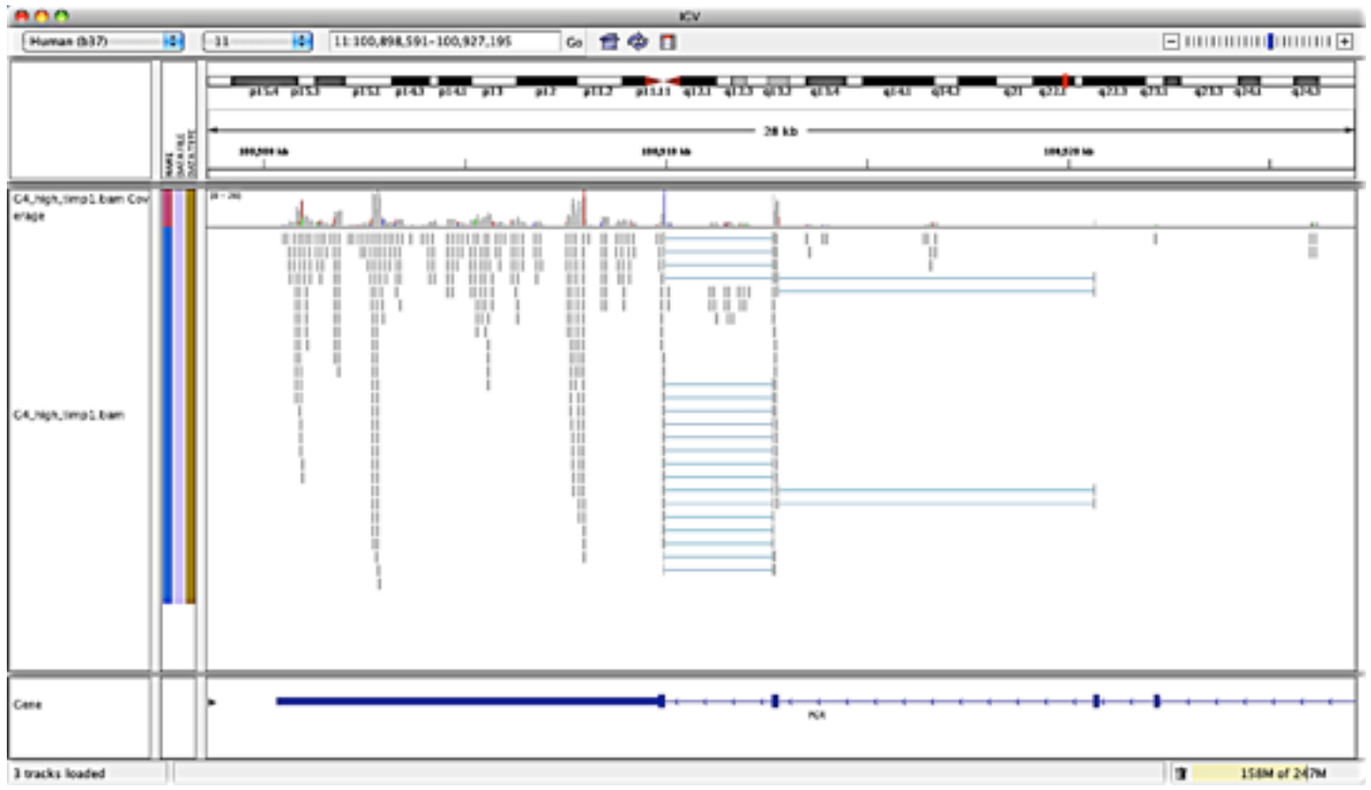

After trimming



# Alignment

- Align reads against a reference genome



# SAM file

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                              ttagctTAGGC
-r001/2                                        CAGCGCCAT
```

The corresponding SAM format is:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

# SAM file

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                              ttagctTAGGC
-r001/2                                        CAGCGCCAT
```

The corresponding SAM format is:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | `[!-?A-~]{1,255}` | Query template NAME |
| 2 | FLAG | Int | $[0,2^{16}-1]$ | bitwise FLAG |
| 3 | RNAME | String | `\*|[!-()+-<>-~][!-~]*` | Reference sequence NAME |
| 4 | POS | Int | $[0,2^{29}-1]$ | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | $[0,2^{8}-1]$ | MAPping Quality |
| 6 | CIGAR | String | `\*|([0-9]+[MIDNSHPX=])+` | CIGAR string |
| 7 | RNEXT | String | `\*|=|[!-()+-<>-~][!-~]*` | Ref. name of the mate/next segment |
| 8 | PNEXT | Int | $[0,2^{29}-1]$ | Position of the mate/next segment |
| 9 | TLEN | Int | $[-2^{29}+1,2^{29}-1]$ | observed Template LENgth |
| 10 | SEQ | String | `\*|[A-Za-z=.]+` | segment SEQuence |
| 11 | QUAL | String | `[!-~]+` | ASCII of Phred-scaled base QUALity+33 |

# SAM file

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                       ATAGCT..............TCAGC
-r003                              ttagctTAGGC
-r001/2                                          CAGCGCCAT
```

The corresponding SAM format is:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | `[!-?A-~]{1,255}` | Query template NAME |
| 2 | FLAG | Int | $[0,2^{16}-1]$ | bitwise FLAG |
| 3 | RNAME | String | `\*|[!-()+-<>-~][!-~]*` | Reference sequence NAME |
| 4 | POS | Int | $[0,2^{29}-1]$ | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | $[0,2^{8}-1]$ | MAPping Quality |
| 6 | CIGAR | String | `\*|([0-9]+[MIDNSHPX=])+` | CIGAR string |
| 7 | RNEXT | String | `\*|=|[!-()+-<>-~][!-~]*` | Ref. name of the mate/next segment |
| 8 | PNEXT | Int | $[0,2^{29}-1]$ | Position of the mate/next segment |
| 9 | TLEN | Int | $[-2^{29}+1,2^{29}-1]$ | observed Template LENgth |
| 10 | SEQ | String | `\*|[A-Za-z=.]+` | segment SEQuence |
| 11 | QUAL | String | `[!-~]+` | ASCII of Phred-scaled base QUALity+33 |

# Alignment - CIGAR strand

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

| Op | BAM | Description |
|---|---|---|
| M | 0 | alignment match (can be a sequence match or mismatch) |
| I | 1 | insertion to the reference |
| D | 2 | deletion from the reference |
| N | 3 | skipped region from the reference |
| S | 4 | soft clipping (clipped sequences present in SEQ) |
| H | 5 | hard clipping (clipped sequences NOT present in SEQ) |
| P | 6 | padding (silent deletion from padded reference) |
| = | 7 | sequence match |
| X | 8 | sequence mismatch |

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

# Alignment - CIGAR strand

```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     gcctaAGCTAA
r004+                   ATAGCT..............TCAGC
r003-                         ttagctTAGGC
r001-                                      CAGCGCCAT

@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

| Op | BAM | Description |
|---|---|---|
| M | 0 | alignment match (can be a sequence match or mismatch) |
| I | 1 | insertion to the reference |
| D | 2 | deletion from the reference |
| N | 3 | skipped region from the reference |
| S | 4 | soft clipping (clipped sequences present in SEQ) |
| H | 5 | hard clipping (clipped sequences NOT present in SEQ) |
| P | 6 | padding (silent deletion from padded reference) |
| = | 7 | sequence match |
| X | 8 | sequence mismatch |

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

# Alignment - CIGAR strand

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | `[!-?A-~]{1,255}` | Query template NAME |
| 2 | FLAG | Int | `[0,2^16-1]` | bitwise FLAG |
| 3 | RNAME | String | `\*|[!-()+-<>-~][!-~]*` | Reference sequence NAME |
| 4 | POS | Int | `[0,2^29-1]` | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | `[0,2^8-1]` | MAPping Quality |
| 6 | CIGAR | String | `\*|([0-9]+[MIDNSHPX=])+` | CIGAR string |
| 7 | RNEXT | String | `\*|=|[!-()+-<>-~][!-~]*` | Ref. name of the mate/next segment |
| 8 | PNEXT | Int | `[0,2^29-1]` | Position of the mate/next segment |
| 9 | TLEN | Int | `[-2^29+1,2^29-1]` | observed Template LENgth |
| 10 | SEQ | String | `\*|[A-Za-z=.]+` | segment SEQuence |
| 11 | QUAL | String | `[!-~]+` | ASCII of Phred-scaled base QUALity+33 |

```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     gcctaAGCTAA
r004+                  ATAGCT..............TCAGC
r003-                        ttagctTAGGC
r001-                                    CAGCGCCAT

@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M * 0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       * 0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    * 0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       * 0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         = 7 -39 CAGCGCCAT          *
```

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

| Op | BAM | Description |
|---|---|---|
| M | 0 | alignment match (can be a sequence match or mismatch) |
| I | 1 | insertion to the reference |
| D | 2 | deletion from the reference |
| N | 3 | skipped region from the reference |
| S | 4 | soft clipping (clipped sequences present in SEQ) |
| H | 5 | hard clipping (clipped sequences NOT present in SEQ) |
| P | 6 | padding (silent deletion from padded reference) |
| = | 7 | sequence match |
| X | 8 | sequence mismatch |

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

# Alignment - Mapping quality

**The Relationship Between Quality Score and Base Call Accuracy**

| Quality Score | Probability of Incorrect Base Call | Inferred Base Call Accuracy |
|---|---|---|
| 10 (Q10) | 1 in 10 | 90% |
| 20 (Q20) | 1 in 100 | 99% |
| 30 (Q30) | 1 in 1000 | 99.9% |

# Alignment - Mapping quality

```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     ~~gcctaAGCTAA
r004+                   ATAGCT..............TCAGC
r003-                         ~~ttagct~~TAGGC
r001-                                        CAGCGCCAT

@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

**The Relationship Between Quality Score and Base Call Accuracy**

| Quality Score | Probability of Incorrect Base Call | Inferred Base Call Accuracy |
|---|---|---|
| 10 (Q10) | 1 in 10 | 90% |
| 20 (Q20) | 1 in 100 | 99% |
| 30 (Q30) | 1 in 1000 | 99.9% |

# Alignment - Mapping quality

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | [!-?A-~]{1,255} | Query template NAME |
| 2 | FLAG | Int | [0,$2^{16}$-1] | bitwise FLAG |
| 3 | RNAME | String | *\|[!-()+-<>-~][!-~]* | Reference sequence NAME |
| 4 | POS | Int | [0,$2^{29}$-1] | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | [0,$2^{8}$-1] | MAPping Quality |
| 6 | CIGAR | String | *\|([0-9]+[MIDNSHPX=])+ | CIGAR string |
| 7 | RNEXT | String | *\|=\|[!-()+-<>-~][!-~]* | Ref. name of the mate/next segment |
| 8 | PNEXT | Int | [0,$2^{29}$-1] | Position of the mate/next segment |
| 9 | TLEN | Int | [-$2^{29}$+1,$2^{29}$-1] | observed Template LENgth |
| 10 | SEQ | String | *\|[A-Za-z=.]+ | segment SEQuence |
| 11 | QUAL | String | [!-~]+ | ASCII of Phred-scaled base QUALity+33 |

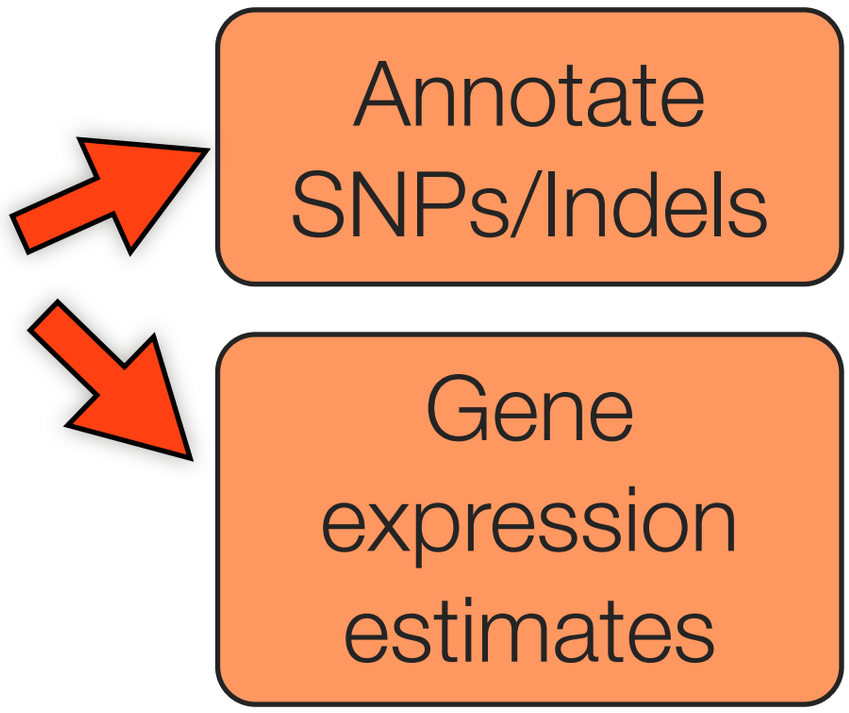
```
coor    12345678901234  5678901234567890123456789012345
ref     AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

r001+         TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     ~~gcctaAGCTAA
r004+                     ATAGCT..............TCAGC
r003-                           ~~ttagct~~TAGGC
r001-                                          CAGCGCCAT

@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M * 0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       * 0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    * 0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       * 0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         = 7 -39 CAGCGCCAT          *
```

**The Relationship Between Quality Score and Base Call Accuracy**

| Quality Score | Probability of Incorrect Base Call | Inferred Base Call Accuracy |
|---|---|---|
| 10 (Q10) | 1 in 10 | 90% |
| 20 (Q20) | 1 in 100 | 99% |
| 30 (Q30) | 1 in 1000 | 99.9% |

# Post-alignment analysis

- Filter out bad aligned reads




Comparative genomics

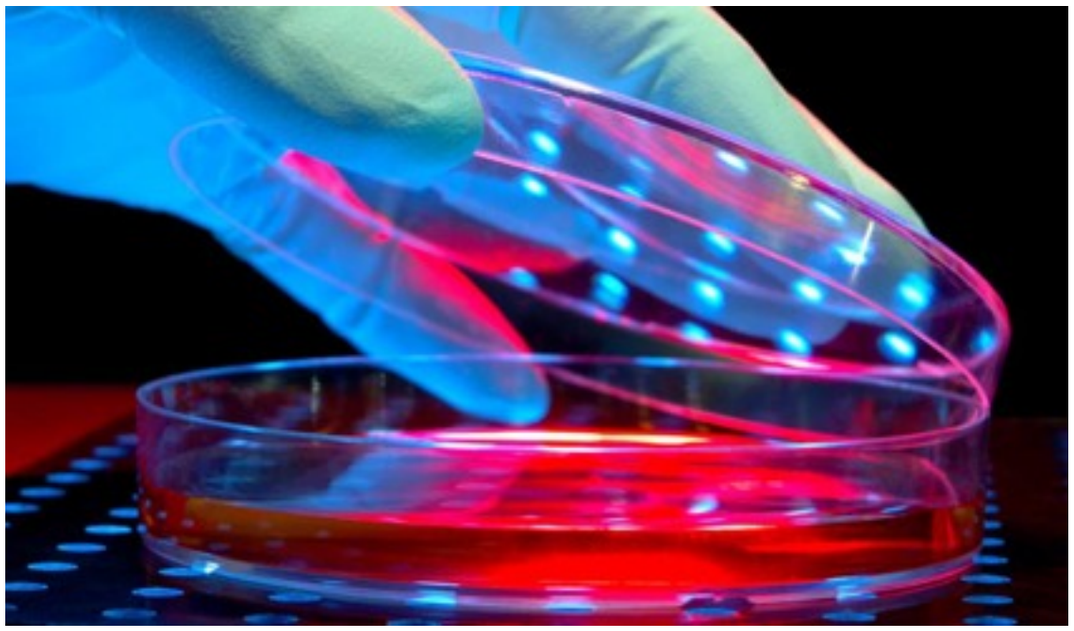
Identify disease-causing variations

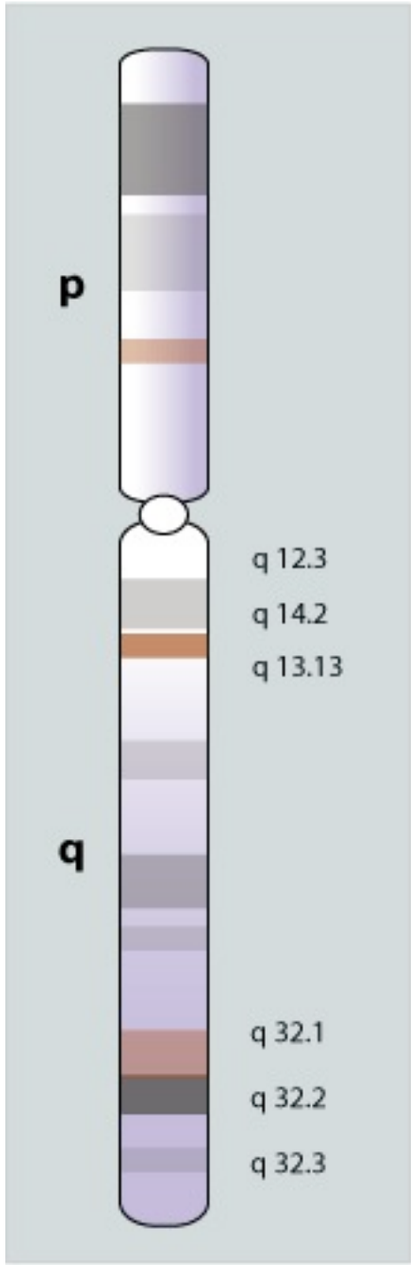
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

# Intro to exercise

# Data

- Exome sequencing data
- Paired-end data - two files
- Breast cancer cell line, chromosome 13






Chromosome 13

# BRCA2

- One of the most common germ line mutation in breast cancer
  - SNP: single nucleotide polymorphism

## A common variant in *BRCA2* is associated with both breast cancer risk and prenatal viability

Catherine S. Healey[1]*, Alison M. Dunning[1]*, M. Dawn Teare[2], Diana Chase[4], Louise Parker[5], John Burn[6], Jenny Chang-Claude[7], Arto Mannermaa[8], Vesa Kataja[9], David G. Huntsman[10], Paul D.P. Pharoah[1], Robert N. Luben[3], Douglas F. Easton[2] & Bruce A.J. Ponder[1]

*These authors contributed equally to this work.

**Inherited mutations in the gene *BRCA2* predispose carriers to early onset breast cancer, but such mutations account for fewer than 2% of all cases in East Anglia. It is likely that low penetrance alleles explain the greater part of inherited susceptibility to breast cancer; polymorphic variants in strongly predisposing genes, such as *BRCA2*, are candidates for this role. BRCA2 is thought to be involved in DNA double strand break-repair[1,2]. Few mice in which *Brca2* is truncated survive to birth; of those that do, most are male, smaller than their normal littermates and have high cancer incidence[3,4]. Here we show that a common human polymorphism (N372H) in exon 10 of *BRCA2* confers an increased risk of breast cancer: the HH homozygotes have a 1.31-fold (95% CI, 1.07–1.61) greater risk than the NN group. Moreover, in normal female controls of all ages there is a significant deficiency of homozygotes compared with that expected from Hardy-Weinberg equilibrium, whereas in males there is an excess of homozygotes: the HH group has an estimated fitness of 0.82 in females and 1.38 in males. Therefore, this variant of BRCA2 appears also to affect fetal survival in a sex-dependent manner.**

In an initial study to investigate whether common *BRCA2* variants alter the risk of breast cancer in the general population, we carried out an association study on six *BRCA2* polymorphisms identified through the BIC database (http://www.nhgri.nih.gov/Intramural_research/Lab_transfer/BIC; Table 1). The genotype distributions of both the exon 10 N372H and the 5′ UTR a-26g polymorphisms approached significant differences between cases and controls in our initial hypothesis-generating study. We thus restricted studies in further case-control series to confirmation of these observations, however, only the N372H findings were confirmed. N372H is the sole *BRCA2* variant resulting in an amino

# BRCA2

- One of the most common germ line mutation in breast cancer
  - SNP: single nucleotide polymorphism

## A common variant in *BRCA2* is associated with both breast cancer risk and prenatal viability

Catherine S. Healey[1]*, Alison M. Dunning[1]*, M. Dawn Teare[2], Diana Chase[4], Louise Parker[5], John Burn[6], Jenny Chang-Claude[7], Arto Mannermaa[8], Vesa Kataja[9], David G. Huntsman[10], Paul D.P. Pharoah[1], Robert N. Luben[3], Douglas F. Easton[2] & Bruce A.J. Ponder[1]

*These authors contributed equally to this work.

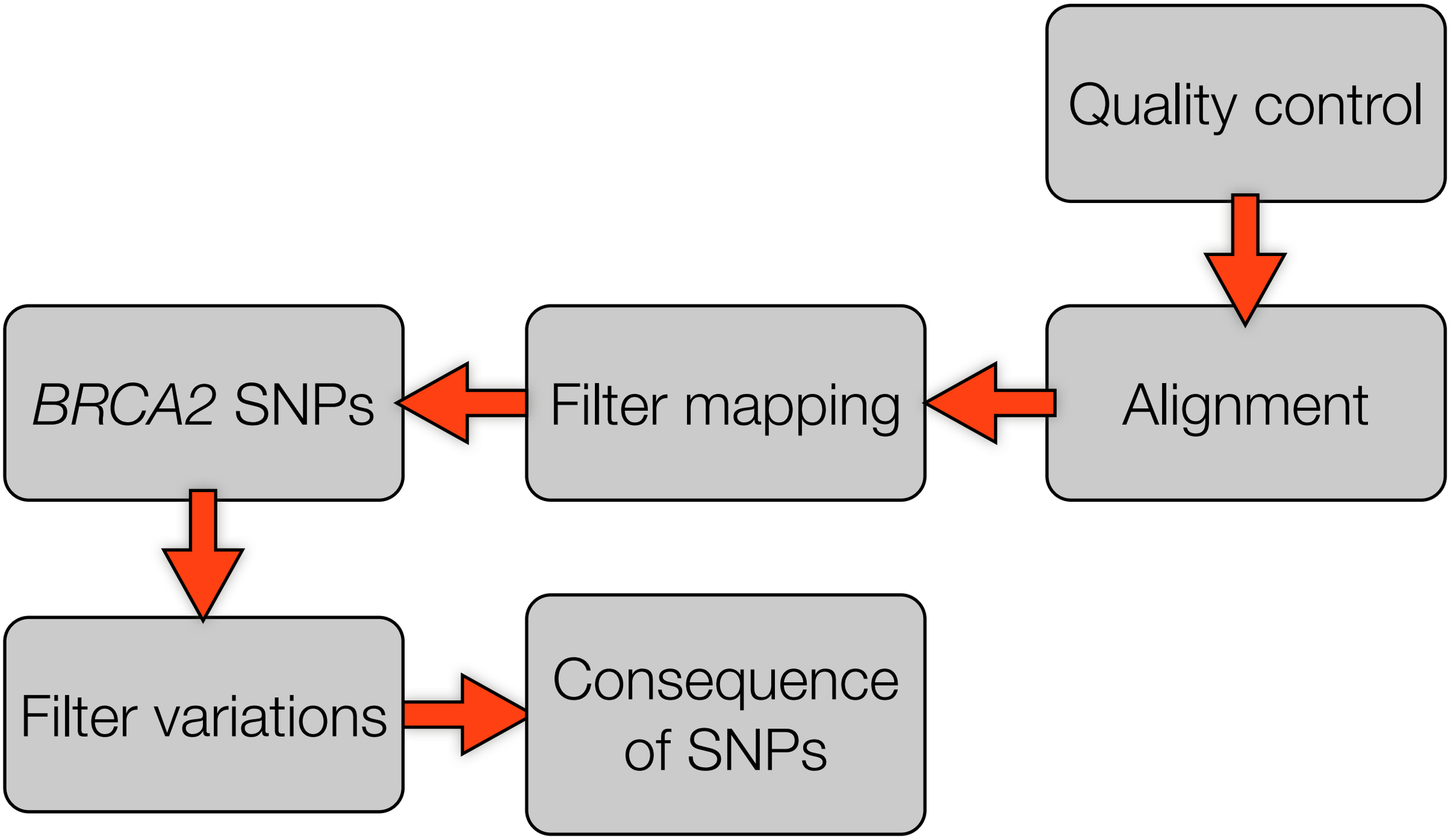
**Inherited mutations in the gene *BRCA2* predispose carriers to early onset breast cancer, but such mutations account for fewer than 2% of all cases in East Anglia. It is likely that low penetrance alleles explain the greater part of inherited susceptibility to breast cancer; polymorphic variants in strongly predisposing genes, such as *BRCA2*, are candidates for this role. BRCA2 is thought to be involved in DNA double strand break-repair[1,2]. Few mice in which *Brca2* is truncated survive to birth; of those that do, most are male, smaller than their normal littermates and have high cancer incidence[3,4]. Here we show that a common human polymorphism (N372H) in exon 10 of *BRCA2* confers an increased risk of breast cancer: the HH homozygotes have a 1.31-fold (95% CI, 1.07–1.61) greater risk than the NN group. Moreover, in normal female controls of all ages there is a significant deficiency of homozygotes compared with that expected from Hardy-Weinberg equilibrium, whereas in males there is an excess of homozygotes: the HH group has an estimated fitness of 0.82 in females and 1.38 in males. Therefore, this variant of BRCA2 appears also to affect fetal survival in a sex-dependent manner.**

In an initial study to investigate whether common *BRCA2* variants alter the risk of breast cancer in the general population, we carried out an association study on six *BRCA2* polymorphisms identified through the BIC database (http://www.nhgri.nih.gov/Intramural_research/Lab_transfer/BIC; Table 1). The genotype distributions of both the exon 10 N372H and the 5′ UTR a-26g polymorphisms approached significant differences between cases and controls in our initial hypothesis-generating study. We thus restricted studies in further case-control series to confirmation of these observations, however, only the N372H findings were confirmed. N372H is the sole *BRCA2* variant resulting in an amino

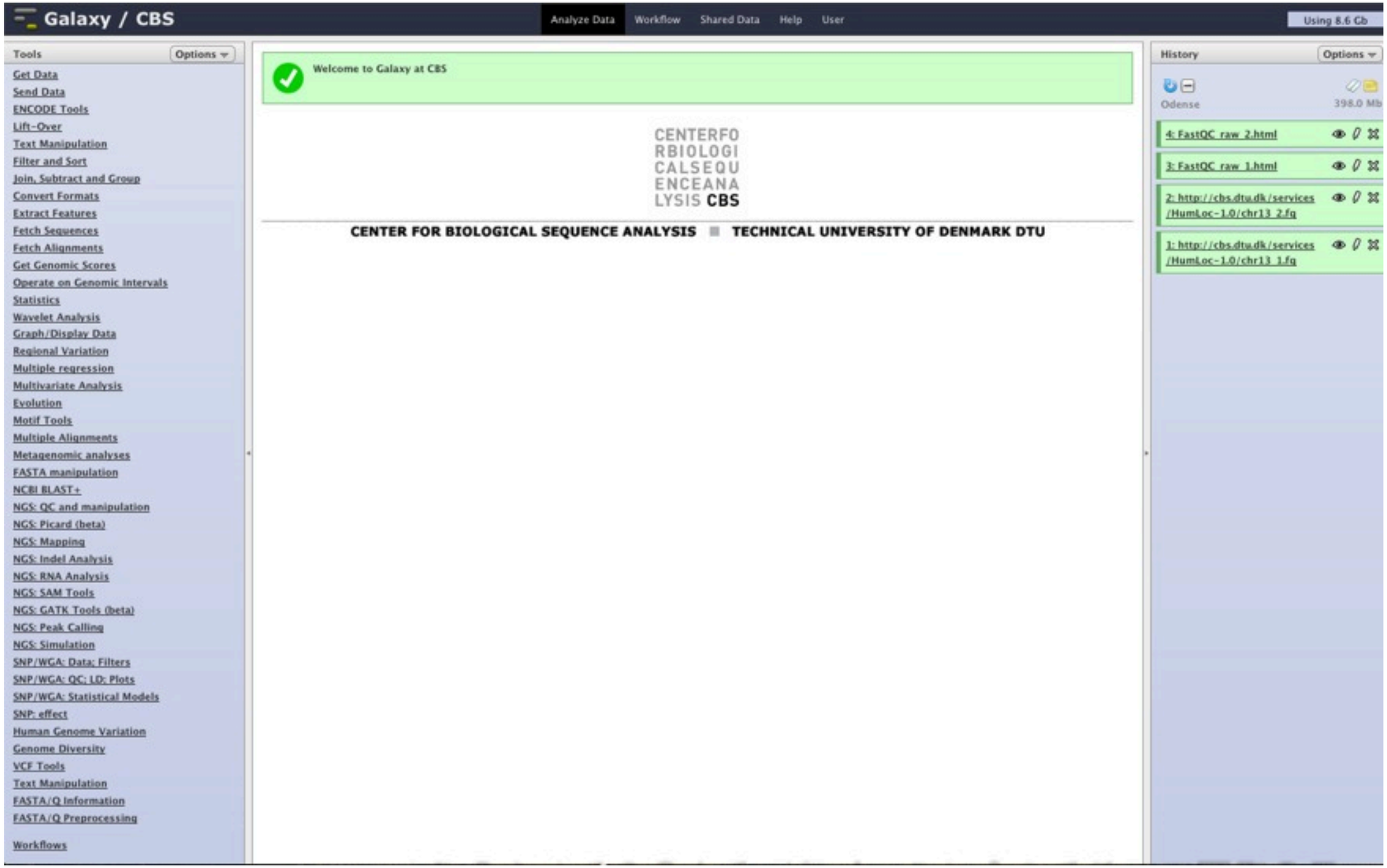
# Purpose of exercise

- Analyze exome sequencing data
- Identify SNPs in *BRCA2* on human chromosome 13
- Identify whether the previous published *BRCA2* SNP N372H is present in the cell line

# Workflow of analysis



# Galaxy


Galaxy / CBS
Analyze Data
Workflow
Shared Data
Help
User
Using 8.6 Gb
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NCBI BLAST+
NGS: QC and manipulation
NGS: Picard (beta)
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis
NGS: SAM Tools
NGS: GATK Tools (beta)
NGS: Peak Calling
NGS: Simulation
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
SNP: effect
Human Genome Variation
Genome Diversity
VCF Tools
Text Manipulation
FASTA/Q Information
FASTA/Q Preprocessing
Workflows
Welcome to Galaxy at CBS
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU
History
Options
Odense
398.0 Mb
4: FastQC raw 2.html
3: FastQC raw 1.html
2: http://cbs.dtu.dk/services/HumLoc-1.0/chr13_2.fq
1: http://cbs.dtu.dk/services/HumLoc-1.0/chr13_1.fq

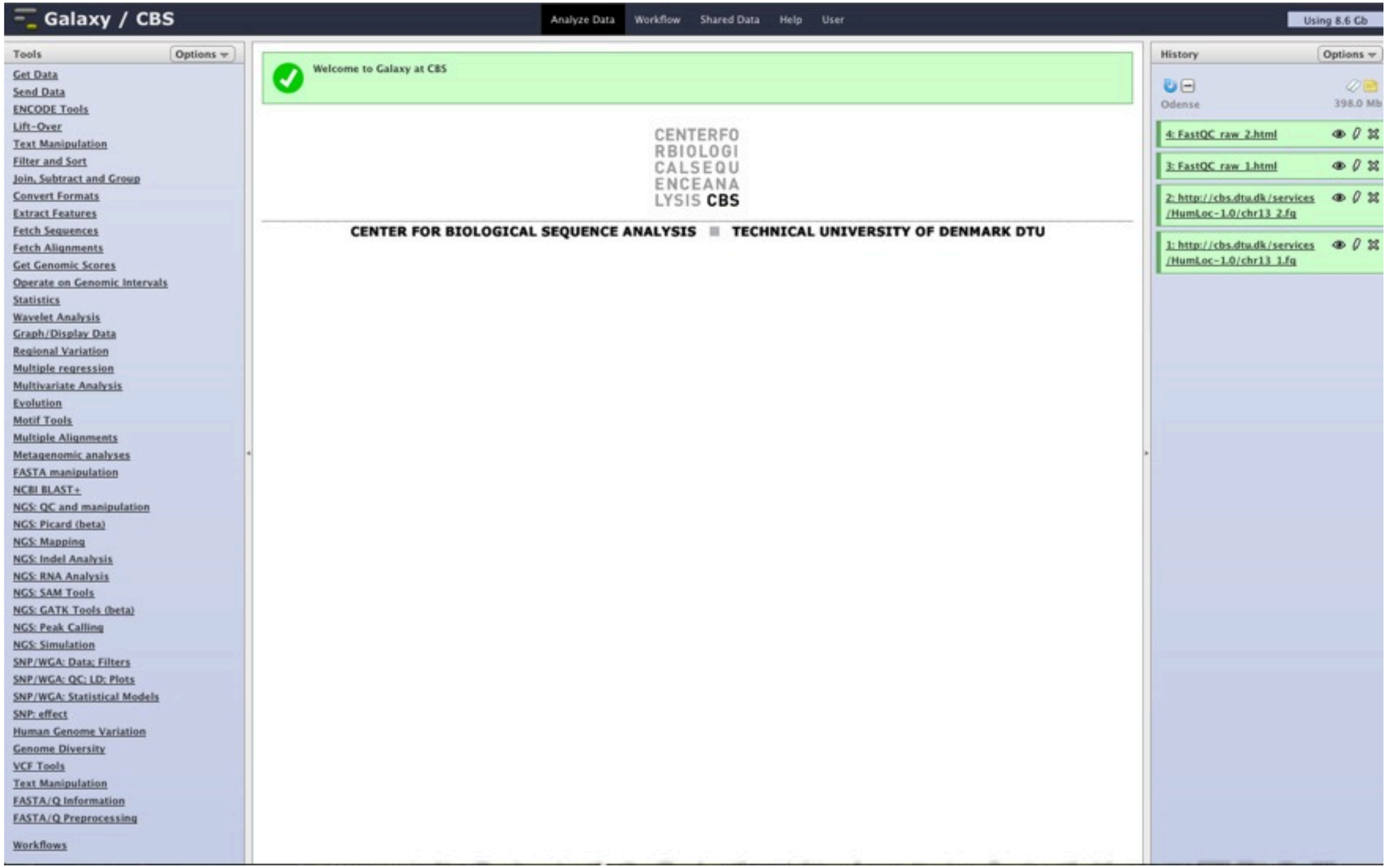
# Galaxy

Job running

Job in queue

Job finished

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



# Galaxy

Job running

Job in queue

Job finished

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

# Your turn to do some work

# Answers

http://wiki.bio.dtu.dk/teaching/index.php/Introduction_to_Systems_Biology_Answers